Introduction

Corpora are principled collections of naturally occurring spoken, written, or multimodal data which are stored in electronic format for quantitative and qualitative analysis. Data from large and representative corpora capture actual language use, and therefore, are reliable resources for list developers to identify the words that second language (L2) learners are likely to encounter in their future use of the target language. Given this benefit, corpora have been widely used to develop and validate word lists for L2 learners, and these lists have made valuable contributions to multiple aspects of L2 vocabulary learning and teaching. This chapter examines current research on corpus-based word lists for L2 learners of English. It begins with critical issues and topics in the area and then proposes some directions for future research.

Critical Issues and Topics

This section is organized into six subsections. The first subsection provides an overview of different kinds of corpus-based word lists. The next four subsections discuss critical issues drawn from the review: the unit of counting, corpus construction, selection criteria, and the reconceptualization of different vocabulary types. The last subsection looks at the application of corpus-based word lists in L2 vocabulary learning and teaching.

Overview of Corpus-Based Word Lists for L2 Learners

Word lists can be classified into lists of single-word items and multiword items. Depending on the learning purposes, each type can be further divided into general service lists and specialized word lists. General service lists focus on the lexical items that are useful for any learning purpose while specialized word lists draw learners’ attention to items that are useful for academic or specific purposes. This section reviews lists of single words and multiword items in turn. Within each type, we will look at general service lists and then specialized word lists.
Lists of Single Words

General service words or high-frequency words (e.g., *know*, *good*) have been widely suggested as the crucial starting point of L2 vocabulary learning. They are small in number, but cover a large proportion of words in different text types. Thus, knowledge of these words may provide learners with a good chance to comprehend language. Given the importance of these words, a large number of general service lists have been developed. Most of them consist of 2,000 to 3,000 items. West’s (1953) General Service List (GSL) is the oldest and most influential list. It has a long-established status and has had a huge impact on L2 vocabulary learning, teaching, and research. However, developed from texts written in the 1930s, the General Service List may not fully reflect current vocabulary. Subsequent studies have addressed this limitation in two ways. The first way is to develop totally new lists which can potentially replace the GSL: Nation’s (2006) BNC2000, Nation’s (2012) BNC/COCA2000, Browne’s (2013) New General Service List, and Brezina and Gablasova’s (2015) New General Service List. The second way is to compile a list with a more manageable size for a specific group of learners from items in existing general service lists: Dang and Webb’s (2016a) Essential Word List.

Specialized word lists can be classified into three types: general academic word lists, discipline-specific word lists, and subject-specific word lists. In this chapter, a discipline refers to a broad group of academic subject areas. For example, academic subject areas such as mathematics and biology can be classified into one disciplinary group (hard sciences) while other subjects such as law and history can be put under another group (soft sciences). The degree of specificity becomes greater from general academic word lists, discipline-specific word lists, to subject-specific word lists.

General academic word lists are concerned with the shared vocabulary among different academic disciplines. They are useful in English for General Academic Purposes programs (see Coxhead, this volume). To date, Coxhead’s (2000) Academic Word List (AWL) is the best-known corpus-based general academic word list. This list was developed with the aim to help L2 learners improve comprehension of academic written text, and has had a great impact on EAP learning and teaching. However, using West’s GSL as the general service vocabulary baseline, the AWL contains a number of current general service words that the GSL fails to capture (Cobb, 2010; Dang & Webb, 2014). Two new lists have been created with the aim to replace the AWL – Gardner and Davies’s (2014) Academic Vocabulary List and Browne, Culligan, and Phillips’s (n.d.) New Academic Word List. Two others were also developed to supplement the AWL by focusing on academic spoken discourse – Nesi’s (2002) Spoken Academic Word List and Dang, Coxhead, and Webb’s (2017) Academic Spoken Word List.

The second type of specialized word list – discipline-specific word lists – narrows its focus on the words that occur across a specific discipline; that is, a group of academic subjects. These lists are suitable for English for Specific Academic Purposes programs. This group includes one written word list – Coxhead and Hirsh’s (2007) EAP Science Word List – and two spoken word lists – Dang’s (2018a) Hard Science Spoken Word List and Dang’s (2018b) Soft Science Spoken Word List.

The third type of specialized word list – subject-specific word lists – have the narrowest focus, but have generated a growing interest from researchers. Subject-specific word lists are also commonly referred to as lists of technical vocabulary (see Liu and Lei in this volume). These lists are relevant to English for Specific Purposes (ESP) programs where all learners plan to study the same subject area (see Coxhead, this volume, for the definition of the
term). They focus on the shared vocabulary in a particular subject area, such as agriculture (Martínez, Beck, & Panza, 2009), applied linguistics (Khani & Tazik, 2013; Vongpumivitch, Huang, & Chang, 2009), business (Konstantakis, 2007), chemistry (Valipouri & Nassaji, 2013), medicine (Hsu, 2013; Lei & Liu, 2016; Wang, Liang, & Ge, 2008), engineering (Hsu, 2014; Ward, 1999, 2009; Watson-Todd, 2017), nursing (Yang, 2015), or environmental science (Liu & Han, 2015). All of these word lists have been based on the language found in written texts.

Lists of Multiword Items

Compared with lists of single words, lists of multiword items are limited in number. Multiword items can refer to any kinds of continuous and discontinuous sequences of words, and how this broad term is interpreted varies from study to study (see Wood, this volume). Most general service lists of multiword items focus on phrasal verbs (Gardner & Davies, 2007; Garnier & Schmitt, 2015; Liu, 2011), collocations (Shin & Nation, 2008), or phrasal expressions (Martinez & Schmitt, 2012). Most specialized multiword lists are general academic word lists. Three of these general academic word lists were derived from written text (Ackerman & Chen, 2013; Byrd & Coxhead, 2010; Durrant, 2009), while the others focused on both spoken and written discourse (Biber, Conrad, & Cortes, 2004; Simpson-Vlach & Ellis, 2010). The focus of these lists also varies from lexical bundles (Biber et al., 2004; Byrd & Coxhead, 2010) to formulas (Simpson-Vlach & Ellis, 2010) to collocations (Ackerman & Chen, 2013; Durrant, 2009).

So far this chapter has briefly reviewed different kinds of word lists, the next sections discuss four critical issues in word list studies: the unit of counting, corpus construction, selection criteria, and the reconceptualization of different vocabulary types.

Unit of Counting

One core issue in word list studies is what should be counted as a word. Different ways of counting words can influence the results of these studies (Nation, 2016). Words can be seen either as single words or multiword items, and the way to count both of these constructs varies between studies.

Table 19.1 shows that four common units of counting in lists of single words are: type, lemma, flemma, and level 6 word family. The word type is the number of unrepeated word forms in a text. For example, the sentence, Counting words is difficult but it is fun, contains eight word forms but seven word types because the word form is is used twice. The lemma (admire) consists of a stem (admire) together with its inflected forms (admires, admired, admiring). Members of a lemma belong to the same word class. The flemma is similar to the lemma, but it does not distinguish parts of speech. For example, form (verb) and form (noun) are counted as two lemmas but only one flemma. The word family consists of a stem, together with its potential inflections and derivations that include affixes up to a certain level in Bauer and Nation’s (1993) scale. This scale is a set of levels (levels 1–7) based on frequency, productivity, predictability, and regularity of affixes. A level 6 word family (admire) would include the stem (admire), and members derived from one or more affix up to level 6 (admirable, admirably, admiration, admirer, admirers, admiringly) (see Nation, 2016 for more details).

Together with the variation in the unit of counting of single words is the continual debate about the best unit for L2 learners. Nation (2016) points out that all of these units are, in fact,
Corpus-Based Word Lists

Table 19.1 The unit of counting in different word lists

<table>
<thead>
<tr>
<th>Kind of word lists</th>
<th>Unit of counting</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Word type</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Discipline-specific word lists</td>
<td>None</td>
</tr>
</tbody>
</table>

* The list is also available in the word family format.
** The list is also available in the flemma format.

different levels of word families in Bauer and Nation’s (1993) scale. Word types should be considered as level 1 word families, lemmas as level 2 families, flemmas as level 2.5, while word families have typically been set at level 6. Therefore, a more accurate question should be which word family level is the most appropriate for a particular group of learners.

The level of word families should match the study purpose (Nation, 2016). Some important questions that need to be answered when determining the unit of counting are (1) who the target users of the list are, (2) whether the list will be used for receptive or productive purposes, and (3) whether the list examines technical words or nontechnical words. To find
the answer to these questions, the learning burden should be taken into account. Learning burden is the amount of effort needed to acquire a word family at a certain level (Nation, 2013). The idea behind the concept of word families is that learners may be able to see that word forms with the same stems are related to each other, and therefore, may find it easier to recognize or learn an unknown word. For example, unhappy is morphologically related to happy rather than a totally unrelated word (cringe).

In terms of list users’ characteristics, the unit of counting should match the characteristics of learners in a particular context. Most existing lists are at either end of Bauer and Nation’s (1993) scale. Research has shown that (1) learners from a certain L1 background had difficulty seeing the relationship between stems and derived forms because their L1 does not have these characteristics (Ward, 2009), and (2) learner knowledge of affixes increases with their vocabulary level (Mochizuki & Aizawa, 2000; Schmitt & Meara, 1997; Schmitt & Zimmerman, 2002). These findings indicate that the most suitable level of counting will vary according to the characteristics of target list users such as their L1 background or their L2 proficiency. Consequently, a certain list may be relevant to a certain group of learners, but may be less applicable to others. In recognition of this need, several recent word lists (Dang, 2018a, 2018b; Dang, Coxhead, & Webb, 2017; Dang & Webb, 2016a; Gardner & Davies, 2014) chose one word family level as the primary unit of counting but also made the lists available in other units of counting so that the lists would be useful to a wide range of learners.

In terms of list purposes, if the list is used for productive purposes, level 1 word families (word types) (Durrant, 2014), level 2 word families (lemmas) (Nation, 2016), or level 2.5 word families (flemmas) are most suitable because knowledge of one word form does not mean that learners are able to use derivations of this word productively (Schmitt & Zimmerman, 2002). However, if the study investigates receptive use, word families at higher levels may be more appropriate (Nation, 2016). Learners with knowledge of the base word or one or two members of the word family may be able to recognize other members of the same word-family when reading or listening. If the study deals with subject-specific technical words, level 1 word families may be most suitable because not all members of a word family are technical words (Chung & Nation, 2004; Nation, 2016).

Let us take Dang and Webb (2016a) Essential Word List (EWL) as an example of how the unit of counting can be selected to match the study purpose. The level 2.5 word family (flemma) was chosen in favor of word families at higher levels for the EWL because it better suits the proficiency level of the target users – L2 beginners, who are still learning the first 1,000 words of English. Because vocabulary size and morphological knowledge are closely related, it is expected that these learners have very limited morphological knowledge, and may not be able to recognize most word family members at higher levels. The level 2.5 word family was chosen over the level 2 word family (lemma) because distinguishing parts of speech may overestimate the learning burden of very closely related items like smile (verb) and smile (noun). It may not be difficult to infer the meaning of smile (noun) in She has a beautiful smile if the meaning of smile (verb) is known.

Similar to single words, there is variation in the way to count multiword items. As shown in the overview, word list studies have examined multiword items from different perspectives: collocations, lexical bundles, phrasal expressions, phrasal verbs, and formulas. Because Wood (this volume) has discussed this issue in detail, this chapter only emphasizes that the variation in the way to interpret multiword items has raised the need for a clear definition of this term in each study. Importantly, a framework that systematically draws all units of counting of multiword items together like Bauer and Nation’s (1993) word family scale
for single words would be useful. Such framework would allow a high degree of consistency in comparing and interpreting results of word list studies.

**Corpus Construction**

The nature of corpora has a great impact on the quality of corpus-based word lists because word selection is primarily based on the information from corpora. This subsection examines the construction of the source corpora (to develop lists) and validating corpora (to validate lists) in turn.

With advances in technology, new lists have been created from larger corpora with the aim to replace existing lists. This, however, does not necessarily mean that these new lists are better if the corpora from which they were developed are not well-designed and clearly described. Source corpora should represent as closely as possible the kind of texts that their target users are likely to encounter often in their language use, and detailed descriptions of their composition should be available to users so that they can judge whether the lists developed from these corpora suit their needs. Biber (1993) provides comprehensive guidelines on how to achieve representativeness for corpora in terms of defining the target population, establishing sampling frames and methods, and evaluating the extent to which the final sample represents the full range of text type variation and linguistic variation. Nation (2016) points out that the construction of corpora for word list studies should carefully consider the content and the size of corpora, and the nature and size of the subcorpora. Unfortunately, while some corpora used to develop word lists satisfy these guidelines, others do not.

Let us take general service lists of single words as an example. These lists aim to capture the words that students are likely to encounter in a range of discourse types. Therefore, an ideal corpus to create a general service list should have equal proportions of spoken and written materials (Nation & Waring, 1997). However, as shown in Table 19.2, only Nation’s (2012) BNC/COCA2000 satisfies this guideline. The BNC/COCA2000 corpus has a good balance between spoken (60%) and written materials (40%) while the corpora used to develop the other lists consist of mainly written materials (75.03%–99.92%). The superiority of the BNC/COCA2000 over the other lists is supported by studies (Dang, 2017; Dang & Webb, 2016b) which compared the five general service lists using lexical coverage, learner vocabulary knowledge, and teacher perceptions of word usefulness as criteria. They found that the BNC/COCA2000 is the most suitable for L2 learners from the perspectives of corpus linguistics, learners, and teachers.

**Table 19.2 Corpora used to develop general service lists of single words**

<table>
<thead>
<tr>
<th>Word lists</th>
<th>Size (words)</th>
<th>Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>West’s (1953) General Service List</td>
<td>5 million</td>
<td>100% 0%</td>
</tr>
<tr>
<td>Nation's (2006) BNC2000</td>
<td>100 million</td>
<td>90% 10%</td>
</tr>
<tr>
<td>Nation’s (2012) BNC/COCA2000</td>
<td>10 million</td>
<td>40% 60%</td>
</tr>
<tr>
<td>Browne’s (2013) NGSL</td>
<td>274 million</td>
<td>75.03% 24.97%</td>
</tr>
<tr>
<td>Brezina and Gablasova’s (2015)</td>
<td>12 billion</td>
<td>99.92% 0.08%</td>
</tr>
<tr>
<td>New General Service List</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Another example is general academic word lists. The aim of these lists is to help learners from a wide range of academic disciplines to comprehend academic written/spoken English. Therefore, the corpora to develop these lists should represent materials from various academic disciplines that these students are likely to encounter frequently in their future study. Importantly, to ensure that the lists are not biased towards the vocabulary in a certain discipline or subject area, the number of subjects per discipline should be the same. So is the number of words per subject. However, as shown in Table 19.3, except for Coxhead (2000) and Dang et al. (2017), corpora used to develop general academic word lists were either unclearly described or unbalanced in terms of the number of subjects per discipline and the number of words per subject.

Apart from the corpora used to create word lists, the corpora used to validate the lists are also important. Testing word lists in independent corpora provides a valid assessment because the results of the validation are not biased towards the corpora from which the lists were derived (Nation & Webb, 2011). Ideally, the validating corpora should be around the same sizes as those used to develop the lists so that they can capture the vocabulary in the target language as well as the source corpus. Unfortunately, the validating stage is absent from most word list studies. Among those having the validation stage, the majority did not meet the mentioned guideline. They either validated word lists in corpora with smaller sizes than those from which the list was developed or did not test the list in corpora of the same or different genres. For example, of the five lists presented in Table 19.3, only Coxhead’s (2000) AWL, Gardner and Davies’s (2014) Academic Vocabulary List and Dang et al. ’s

Table 19.3 Corpora used to develop general academic word lists

<table>
<thead>
<tr>
<th>Word lists</th>
<th>Developing corpus</th>
<th>Components</th>
<th>Text types</th>
<th>Size</th>
<th>Subcorpus</th>
<th>Number of disciplines</th>
<th>Words/discipline</th>
<th>Number of subjects/discipline</th>
<th>Words/subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coxhead (2000) AWL</td>
<td>100% written</td>
<td>University textbooks, articles, book chapters, laboratory manuals</td>
<td>3.5 million</td>
<td>4</td>
<td>875,000</td>
<td>7</td>
<td>125,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Browne et al. (n.d.) NAWL</td>
<td>98.90% written</td>
<td>Journal articles, nonfictions, student essays, lectures, seminars</td>
<td>288 million</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nesi (2002) SAWL</td>
<td>100% spoken</td>
<td>Lectures, seminars</td>
<td>1.6 million</td>
<td>4</td>
<td>Around 400,000</td>
<td>10–19</td>
<td>2,038–36,251</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gardner and Davies (2014) AVL</td>
<td>100% written</td>
<td>Journal articles, newspapers, magazines</td>
<td>120 million</td>
<td>Not applicable</td>
<td>Not applicable</td>
<td>9</td>
<td>8 to 22 million</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dang, Coxhead, and Webb (2017) ASWL</td>
<td>100% spoken</td>
<td>Lectures, seminar, labs, tutorials</td>
<td>13 million</td>
<td>4</td>
<td>3.5 million</td>
<td>6</td>
<td>500,000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Corpus-Based Word Lists

(2017) Academic Spoken Word List were tested against independent academic and non-academic corpora. Only the validating corpora of the Academic Spoken Word List had a similar size as the corpus used to develop the list (13 million words). The other validating corpora were much smaller than the source corpora – 678,000 words compared to 3.5 million words (Academic Word List) and 16 million to 83 million words compared to 120 million words (Academic Vocabulary List).

Selection Criteria

The third issue to consider when evaluating corpus-based word list studies is the selection criteria. Most studies (e.g., Brezina & Gablasova, 2015; Byrd & Coxhead, 2010) rely solely on objective corpus-driven criteria. Common selection criteria in both lists of single words and multiword items are frequency, range, and dispersion. Frequency is the number of occurrences of a word or a multiword in the entire corpus. Range indicates the number of different texts or subcorpora that the word/multiword occurs in. Dispersion shows how evenly a word/multiword is distributed across different texts or subcorpora. The higher frequency, wider range, and more even distribution a word/multiword has, the more likely learners are to encounter it in their language use. Apart from frequency, range, and dispersion, some criteria are unique to the selection of single words or multiword items. Lexical coverage – the percentage of known words in a corpus – is a common criterion to select single words. The higher the lexical coverage a word provides, the greater value it may have to learners. N-grams, Mutual Information, t-scores, and log-likelihood are popular criteria to select multiword items. N-grams refer to the length of the word sequence while the other criteria indicate the extent to which the items in a certain sequence occur more frequently together than would be expected by chance.

Relying solely on objective corpus-driven criteria to select and validate word lists has two strengths (Nation, 2016). First, it results in a list that is replicable, which then allows the comparison of the list with other lists using different corpora or different criteria. Second, this approach makes it easier to divide the list into frequency-based sub-lists that can be set as short-term learning goals. However, lists developed from this approach will inevitably be affected by the nature of the corpus from which they were developed. Consequently, some items that may be useful for L2 learners may be absent from these lists because they do not meet the objective selection criteria of the corpus from which the items were derived.

Two approaches have been taken to address this limitation. The first approach is to use objective corpus-driven information as the main criterion in word selection but also use subjective criteria as a guide to adjust these objective criteria so that the lists are more suitable and useful for a particular group of learners. Let us take Dang et al.’s (2017) Academic Spoken Word List as an example. This list is aimed at L2 learners in English for General Academic Purposes programs who plan to study different academic disciplines and have different language proficiency levels. Range, frequency, and dispersion were used as the primary criteria to select items in this list. However, given the limited learning time and slow vocabulary growth rates of L2 learners compared to L1 children, different versions adopting different range, frequency, and dispersion cutoff points were compared. Four key considerations – (1) size and coverage, (2) word families outside general high frequency word-families, (3) distribution across the four disciplinary subcorpora, and (4) adaptability to learners’ levels – were used as the guide to see which version would have the greatest pedagogical value. The first two criteria ensured that the Academic Spoken Word List contained a smaller number of items but provided higher lexical coverage than a general service
list. These criteria were necessary because if the Academic Spoken Word List either had a larger size but provided less coverage than a general service list, or mainly consisted of general service words, it would not have much value for the target users. It would simply be more useful to learn with existing general service lists. The last two criteria were important because they made sure that the Academic Spoken Word List could benefit the target users irrespective of their academic disciplines and proficiency levels.

Another example is Shin and Nation’s (2008) list of high-frequency collocations of spoken English. This list targets elementary learners of English whose vocabulary level is at the most frequent 1,000 words. Frequency of word sequences was used as the main criterion for selection. However, Shin and Nation also made some subjective judgments by only selecting collocations with node words that were content words and among the most frequent 1,000 words. This decision ensured that the selected collocations were meaningful units that would have value for teaching and deliberate learning. Thus, by learning items from the list, it would help strengthen and enrich knowledge of known words. Moreover, Shin and Nation only selected collocations that met a certain frequency threshold and expressed a complete meaning so that their list was a manageable size and useful for the users.

The second approach to selecting items is to use subjective criteria together with objective corpus-driven criteria in word list development and validation. Let us look at some examples of this approach. West (1953) used one objective corpus-driven criterion (frequency) as the primary criterion to select items for his GSL but also used other subjective criteria (ease of learning, necessity, cover, stylistic level, and emotional neutrality) so that the list was suitable for L2 learners. Similarly, range, frequency, and dispersion were the main criteria to select Nation’s (2006) BNC2000 and Nation’s (2012) BNC/COCA2000 words. However, items that did not meet the objective criteria but were considered useful for L2 learners such as informal spoken words (e.g., ok, hello), modern words (e.g., internet, web), survival words (e.g., delicious, excuse), numbers, weekdays, and months were also included so that these lists were suitable for L2 learning and teaching. Similarly, Ackerman and Chen (2013) used objective corpus-driven criteria (frequency, range, N-grams, MI, t-score, word classes) as the primary criteria to select items for their Academic Collocation List, and then used subjective criterion (experts’ judgment) to filter items in the list.

Using corpus-driven information together with subjective criteria may lead to the development of word lists that better serve L2 learning and teaching purposes compared to solely relying on objective corpus-driven information to develop word lists. This claim is supported by Dang, Webb, and Coxhead’s (under review) study, which examined the relationships between lexical coverage, L2 learner knowledge, and English language teacher perceptions of the usefulness of general service words. These researchers found that although these three factors were related, the relationships between lexical coverage and the other two factors were not as strong as that between learner vocabulary knowledge and teacher perceptions. This finding indicates that, although corpora can provide valuable information for word list construction, the use of subjective criteria from learners and teachers might be also be necessary to make corpus-based word lists more relevant to L2 learning and teaching purposes.

Reconceptualization of Different Kinds of Vocabulary

One issue that arises from the review of word list studies is the reconceptualization of different kinds of words. According to Nation (2001), words can be divided into different 1,000-item bands based on their frequency and range. Items at the first 1,000-word band are
the most frequent and wide ranging words while those at the second 1,000-word band have lower frequency and narrower range. The further the 1,000-word bands are from the first 1,000-word band, the less frequent these bands are. Based on this, Nation (2001) categorized vocabulary into four types – high-frequency words, low-frequency words, academic words, and technical words. High-frequency words are those at the first and second 1,000-word bands. Academic words and technical words are considered as specialized vocabulary and have lower frequency compared to high-frequency words. Low-frequency words are those outside high-frequency words, academic words, and technical words. This traditional classification has been widely accepted, and has had a great influence on word list studies. However, recent research has questioned this classification regarding the boundary between high and low-frequency words, and between general and specialized words. Such questions have been reflected in word list studies.

Regarding the boundary between high- and low-frequency words, 2,000 items has been widely accepted as the number of high-frequency words. As a result, most general service lists (Brezina & Gablasova, 2015; Nation, 2006, Nation, 2012; West, 1953) have around this size. However, investigating the issue from different perspectives (frequency and incidental acquisition, frequency and use of graded readers, and lexicography and dictionary defining vocabulary), Schmitt and Schmitt (2014) suggest that this cutoff point should be expanded to 3,000 (1st–3rd 1,000) words, and items outside general high-frequency words should be divided into mid-frequency words (4th–9th 1,000) and low-frequency words (beyond 9th 1,000). Recognizing the limitation of Nation’s (2001) classification, Nation (2013) then reconstructed his classification by adding the mid-frequency word level. Compared with the traditional classification, the classification of high-, mid-, and low-frequency words provides a more precise learning goal for L2 learners at different proficiency stages (see Vilkaitė-Lozdiena and Schmitt, this volume).

Regarding the boundary between general and specialized vocabulary, following Nation’s (2001) classification, the most common approach towards developing specialized word lists (e.g., Browne et al., n.d.; Coxhead, 2000; Coxhead & Hirsh, 2007; Nesi, 2002) has been to consider academic and technical vocabulary as being outside of general service vocabulary, and therefore, general service words are not included in these lists. For example, Coxhead’s (2000) Academic Word List does not include items from West’s (1953) list. The strength of this approach is that it considers learners’ existing knowledge of general vocabulary, and enables learners and teachers to avoid repeatedly learning and teaching known items. However, specialized word lists following this approach are inevitably affected by the nature of the general service lists on which they were built.

In recognition of this limitation, a more recent approach has been to consider specialized vocabulary as a separate kind of vocabulary that does not directly relate to general service vocabulary (e.g., Gardner & Davies, 2014; Lei & Liu, 2016; Ward, 1999, 2009). For instance, Gardner and Davies (2014) included in their Academic Vocabulary List all items that have wider range and higher frequency in academic text than non-academic text that meet the selection criteria even if they are included in existing general service lists. Specialized word lists following this approach cut across the high-, mid-, and low-frequency levels, and are not affected by the limitation related to existing general service lists. Yet, as learners’ knowledge of general vocabulary is not considered in the list development, there may be a risk of repeatedly teaching and learning known items, which may result in inefficient learning and teaching time.

The third approach (Dang et al., 2017; Dang, 2018a, 2018b) expands on the two previous approaches. It views specialized vocabulary as a separate kind of vocabulary but also looks
at it in relation to general vocabulary. Let us take Dang et al.’s (2017) Academic Spoken Word List as an example. Following the second approach, general service words were still included in the list if they met the selection criteria. Following the first approach, learners’ knowledge of general vocabulary was also considered in the list development. However, instead of setting a fixed benchmark of the number of words these learners need to know before learning items from the list, the Academic Spoken Word List was divided into four levels according to Nation’s (2012) BNC/COCA lists. Depending on their current level of general vocabulary, learners can skip certain levels of the Academic Spoken Word List. This approach proposes a systematic way to integrate specialized vocabulary and general vocabulary. On one hand, it reflects the blurred boundary between general and specialized vocabulary. On the other hand, it ensures that learners’ knowledge of general vocabulary is taken into account.

Corpus-Based Word Lists in L2 Learning and Teaching

How to apply corpus-based word lists in L2 curriculum and instruction is a growing concern of current vocabulary research. While there is still a great deal of interest in replacing existing lists with new lists by making use of larger corpora and more powerful statistical measures, an interesting and meaningful move in corpus-based word list studies is to make existing lists more suitable to a particular learning and teaching context.

One trend is to create smaller lists that match the proficiency level of a particular group of learners. Ward’s (2009) Basic Engineering English word list has only 299 items, but may allow its target users (low proficiency foundation engineering undergraduates in Thailand) to recognize around 16% of the words in their textbooks. Dang and Webb’s (2016a) Essential Word List consists of the best 800 items (in terms of lexical coverage) from West’s GSL, Nation’s (2006) BNC2000, Nation’s (2012) BNC/COCA2000, and Brezina and Gablasova’s (2015) New General Service List. Knowledge of the Essential Word List headwords and lemmas may enable its target users (L2 beginners) to recognize 60% and 75% of the words in a wide range of spoken and written English texts, respectively.

The second trend is to make a list that is adaptable to learners’ proficiency levels (Dang, 2018a, 2018b; Dang et al., 2017) or break a list into sub-lists with manageable size to fit in individual courses within a language program (Coxhead, 2000; Dang, 2018a, 2018b; Dang et al., 2017; Dang & Webb, 2016a). For instance, as discussed previously, Dang et al.’s Academic Spoken Word List is divided into four levels based on lists of general vocabulary, and learners can focus their attention on the Academic Spoken Word List levels that are relevant to their current level of general vocabulary. Each level is further divided into sub-lists with manageable sizes to allow easy incorporation of the list in language learning programs.

The third trend is to help learners achieve deeper knowledge of items in existing lists. Shin and Nation’s (2008) collocation list allows L2 beginners to expand their knowledge of the most frequent 1,000 words from single words to multiword items. Garnier and Schmitt’s (2015) PHaVE List supports learners to acquire the most frequent meanings of items in Liu’s (2011) phrasal verb list.

Let us now look at the contributions of corpus-based word lists to different aspects of L2 vocabulary learning and teaching. The most obvious value of these lists can be seen in the area of setting learning goals. Different kinds of words in English do not occur with similar frequencies. Therefore, determining what and how many words need to be learned
at different stages of language learning is extremely important. It ensures that learners can get the best return for their learning effort. Word lists provide course designers with a practical way to decide which words to focus on in a course. The new classification of vocabulary into three levels of high-, mid-, and low-frequency words provides more precise long-term learning goals for L2 learners, and better supports their continual vocabulary development. Similarly, the reconceptualization of specialized vocabulary provides EAP/ESP course designers with more flexibility when integrating different kinds of word lists into the curriculum to meet learners’ target subject areas, language proficiency, and learning purposes (see Dang, 2018a for a model of selecting relevant word lists for different language programs).

The value of corpus-based word lists has not just been restricted to setting learning goals but also includes the development of vocabulary tests. Tests developed based on corpus-based word lists such as Vocabulary Levels Test (Nation, 1990, Schmitt, Schmitt, & Clapham, 2001; Webb, Sasao, & Ballance, 2017) and Vocabulary Size Test (Nation & Beglar, 2007) have been widely used to assist teachers in setting learning goals and monitoring learners’ progress.

With respect to L2 learning and teaching, discussion of word lists is commonly associated with negative ideas of deliberate rote learning of words out of context. However, it is worth stressing that when effectively done with the right words, such learning can positively contribute to L2 vocabulary development (see Nakata, this volume). Well-designed word lists, therefore, are essential resources for effective deliberate learning. Importantly, presenting items that learners should learn in the format of word lists does not mean that these lists should be learned and taught solely through decontextualized methods (Coxhead, 2000; Nation, 2016). Deliberate learning should account for no more than 25% of the amount of time of a well-balanced vocabulary learning program (Nation, 2007, 2013; Webb & Nation, 2017). The remaining time should be equally spent in learning from meaning-focused input, meaning-focused output, and fluency development. Learning from meaning-focused input means that learners acquire new words through listening or reading activities while learning from meaning-focused output means that learners acquire new words through speaking and writing activities. Fluency development involves all four skills and does not aim to help students learn new words but to make them become more fluent at using known words. In all three cases, the main focus of attention is on meaning; that is, to understand the messages that are received or to produce meaningful messages. These strands are only effective if learners have sufficient vocabulary knowledge to make them truly meaning-focused. This means that for learning from meaning-focused input and output to happen, learners should already know at least 95% of the words in spoken materials (van Zeeland & Schmitt, 2013) and 98% of the words in written materials (Hu & Nation, 2000). The low proportion of words beyond learners’ current vocabulary knowledge allows them to pay more attention to new words and learn them from contexts or with reference to dictionaries or glossaries. Similarly, materials for fluency development activities should involve no or very few unknown words so that learners can process and produce the known words at a fast speed. For these reasons, corpus-based word lists play a key role in the selection and development of learning materials for meaning-focused input, meaning-focused output, and fluency development activities. Using these lists together with vocabulary processing programs like RANGE (Heatley, Nation, & Coxhead, 2002), AntwordProfiler (Anthony, n.d), or Lextutor (Cobb, n.d) allows teachers to analyze vocabulary in the materials, and then adapt it so that these materials are relevant to learners’ vocabulary levels (Webb & Nation, 2008).
Future Directions

There are two directions for future research on corpus-based word lists. The first direction is to focus more on some under-researched areas of corpus-based word lists, including mid-frequency vocabulary, spoken vocabulary, subject-specific vocabulary, and multiword units. The second direction is to bring more pedagogical value to corpus-based research on word lists. Most corpus-based word lists rely solely on objective corpus-driven criteria; therefore, future research should also include other sources of information such as teacher perceptions and learner vocabulary knowledge to support the information from corpora.

Further Reading


This article includes discussion of important issues in this chapter: unit of counting, corpus construction, selection criteria, and reconceptualization of general academic vocabulary.


This article provides comprehensive discussion of high-frequency word lists and how to evaluate lists of different units of counting.


This book provides very detailed guidelines for making word lists for language teaching and testing. It covers important issues in word lists studies such as identifying the list purposes, deciding the unit of counting, constructing corpora, and criteria to select words.


Chapter 8 of this book describes the basic steps of making word lists.


Chapter 1 of this book explains the definition of corpora, basic steps of making a corpus, and the most common techniques of analyzing language in a corpus.

Related Topics

The different aspects of vocabulary knowledge, defining multiword items, high-, mid-, and low-frequency words, academic vocabulary, technical vocabulary

References


