Steps for Creating a Specialized Corpus and Developing an Annotated Frequency-Based Vocabulary List

Marie-Claude Toriida

This article provides introductory, step-by-step explanations of how to make a specialized corpus and an annotated frequency-based vocabulary list. One of my objectives is to help teachers, instructors, program administrators, and graduate students with little experience in this field be able to do so using free resources. Instructions are first given on how to create a specialized corpus. The steps involved in developing an annotated frequency-based vocabulary list focusing on the specific word usage in that corpus will then be explained. The examples are drawn from a project developed in an English for Academic Purposes Nursing Foundations Program at a university in the Middle East. Finally, a brief description of how these vocabulary lists were used in the classroom is given. It is hoped that the explanations provided will serve to open the door to the field of corpus linguistics.

Cet article présente des explications, étape par étape, visant la création d'un corpus spécialisé et d'un lexique annoté et basé sur la fréquence. Un de mes objectifs consiste à aider les enseignants, les administrateurs de programme et les étudiants aux études supérieures avec peu d'expérience dans ce domaine à réussir ce projet en utilisant des ressources gratuites. D'abord, des directives expliquent la création d'un corpus spécialisé. Ensuite, sont présentées les étapes du développement d'un lexique visant le corpus, annoté et basé sur la fréquence. Les exemples sont tirés d'un projet développé dans une université du Moyen-Orient pour un cours d'anglais académique dans un programme de fondements de la pratique infirmière. En dernier lieu, je présente une courte description de l'emploi en classe de ces listes de vocabulaire. J'espère que ces explications ouvriront la porte au domaine de la linguistique de corpus.

KEYWORDS: corpus development, specialized corpus, nursing corpus, spaced repetition, Antconc

A corpus has been defined as "a collection of sampled texts, written or spoken, in machine readable form which may be annotated with various forms of linguistic information" (McEnery, Xiao, & Tono, 2006, p. 6). One area of research in corpus linguistics has focused on looking at the frequency of the words used in real-world contexts. Teachers have used such information for the purpose of increasing language learner success. For example, the seminal

http://dx.doi.org/1018806/tesl.v34i1.1255

General Service List (GSL; West, 1953), a list of approximately 2,200 words, was long said to represent the most common headwords of English, as they comprise, or *cover*, approximately 75–80% of all written texts (Nation & Waring, 1997) and up to 95% of spoken English (Adolphs & Schmitt, 2003, 2004). Similarly, the Academic Word List (AWL; Coxhead, 2000) is a 570-word list of high-frequency word families, excluding GSL words, found in a variety of academic texts. It has been shown to cover approximately 10% of a variety of textbooks taken from different fields (Coxhead, 2011). Thus, the *lexical coverage* of the GSL and AWL combined is between 85% and 90% of academic texts (Neufeld & Billuroğlu, 2005).

More recent versions of these classic lists include the New General Service List (new-GSL; Brezina & Gablasova, 2015), the New General Service List (NGSL; Browne, Culligan, & Phillips, 2013b), the New Academic Word List (NAWL; Browne, Culligan, & Phillips, 2013a), and the Academic Vocabulary List (AVL; Gardner & Davies, 2014). Large corpora of English also exist, such as the recently updated *Corpus of Contemporary American English* (Davies, 2008–) and the *British National Corpus* (2007). These corpora are based on large amounts of authentic texts from a variety of fields.

Hyland and Tse (2007), however, noted that many words have different meanings and uses in different fields, hence the need to learn context-specific meanings and uses. They further stated as a criticism of the AWL, "As teachers, we have to recognize that students in different fields will require different ways of using language, so we cannot depend on a list of academic vocabulary" (p. 249). As a means to address this concern, specialized corpora specific to particular fields and contexts have been developed in recent years. For examples of academic nursing corpora, see Budgell, Miyazaki, O'Brien, Perkins, and Tanaka (2007), and Yang (2015).

Nursing Corpus Project: Context and Rationale

Our institution, located in the Middle East, offers two nursing degrees: a Bachelor of Nursing degree and a Master of Nursing degree. The English for Academic Purposes (EAP) Nursing Foundations Program is a one-year, three-tiered program. It has the mandate to best prepare students for their first year in the Bachelor of Nursing program. Our students come from a variety of educational and cultural backgrounds. Some students are just out of high school, while others have been practicing nurses for many years. We felt that a corpus-based approach for targeted vocabulary learning would best serve our diverse student population, and be an efficient way to address the individual linguistic gaps hindering their ability to comprehend authentic materials used in the nursing program (Shimoda, Toriida, & Kay, 2016).

One factor that greatly affects reading comprehension is vocabulary knowledge (Bin Baki & Kameli, 2013). Reading comprehension research has shown that the more vocabulary is known by the reader, the better their reading comprehension will be. For example, Schmitt, Jiang, and Grabe (2011) found a linear relationship between the two. Previous researchers also looked at this relationship in terms of a vocabulary knowledge threshold for successful reading comprehension. Laufer (1989) claimed that knowledge of 95% of the words in a text was needed for minimal comprehension in an academic setting, set as an achievement score of 55%. In a later study, Hu and Nation (2000) suggested that 98% lexical coverage of a text was needed for adequate comprehension when reading independently, with no assistance from a gloss or dictionary. One problem raised was how to define "adequate comprehension." Laufer and Ravenhorst-Kalovski (2010) later suggested that 95% vocabulary knowledge would yield adequate comprehension if adequate comprehension was defined as "reading with some guidance and help" (p. 25). They further supported Hu and Nation's (2000) findings that 98% lexical knowledge was needed for unassisted independent reading. These findings highlight the importance of vocabulary knowledge in reducing the reading burden. This is especially critical when dealing with second language learners who are expected to read nursing textbooks high in academic and technical vocabulary.

To best facilitate the transition from the EAP to the nursing program, a corpus was thus developed from an introductory nursing textbook intensively used in the first-year nursing courses at our institution. From this corpus of 152,642 *tokens* (total number of words), annotated vocabulary lists based on word frequency were developed for the first 2,500 words of the corpus (25 lists of 100 words), as they constituted close to 95% of the text. The lists included, for each word, the part(s) of speech, a context-specific definition, high-frequency collocation(s), and a simplified sample sentence taken from the corpus. An individual vocabulary acquisition program using these lists was later introduced at all levels of the EAP program.

The teacher participants involved in this project had no prior experience developing a corpus. Compiling a corpus from a textbook was a long and extensive task, one that preferably should be done as a team. To get acquainted with the process, teachers may want to try developing a corpus and annotated frequency-based vocabulary lists from smaller, more specific sources to fit their specific needs, such as graded readers, novels, journal articles, or textbook chapters. One advantage of doing this is statistically knowing the frequency of the words that compose the corpus and how they are used in that specific context. This can validate intuition and facilitate the selection of key vocabulary or expressions to be taught and tested. Similarly, it can help make informed decisions as to what words might be best presented in a gloss. Another advantage is being able to extract high-frequency collocations specific to the target corpus. In short, a corpus-based approach is a form of evidence-based language pedagogy that provides teachers with information to guide decisions regarding vocabulary teaching, learning, and testing. It is important to note, however, that the smaller the number of words in a corpus, the lower its stability, reliability, and generalizability (Browne, personal communication, March 12, 2013). Having said that, a smaller corpus can still be of value for your teaching and learning goals. As Nelson (2010) noted, "the purpose to which the corpus is ultimately put is a critical factor in deciding its size" (p. 54).

This article will provide a practical explanation of the steps involved in creating a specialized corpus and frequency-based vocabulary list using free resources. Suggestions will also be presented on how to annotate such a list for student use. Finally, a brief explanation of how annotated lists were used in our EAP program will be given. The following is intended as an introductory, step-by-step, practical guide for teachers interested in creating a corpus.

Preparing a Corpus

Target Materials

The first important step in creating a corpus is thinking about your teaching context, your students' language needs, and how the corpus will be used. This will help determine what materials the corpus will comprise. Materials could include a textbook or textbook chapter, a collection of journal articles, a novel, graded readers, course materials, or a movie script, among other texts. Once this is decided, the materials need to be converted into a word processing document. Electronic copies of books may be available through your institution's library. When only hard copies or PDF files are available, some added steps are necessary. Hard copies should first be scanned and saved as PDF files. Optical character recognition (OCR) software can then be used. Many online OCR programs will allow you to convert a limited number of documents or pages for free, such as *Online OCR* (www.onlineocr.net). Another option is to use *AntfileConverter* (Anthony, 2015), freely available software (with no page limits) that converts PDF files to plain text (txt) format, which can then be cut and pasted into a word processing document. A final option is to purchase OCR software. Check with your institution's IT department, as they may have OCR software available. Documents converted through OCR software require a final check against the original as the conversions are not always 100% accurate.

Word Elimination

Word elimination refers to the process of deleting words from the corpus that are not considered content words. This is done to prepare the corpus for analysis. Reference sections and citations can first be deleted. Repetitive textbook headings, figure and table headings, proper nouns, and names of institutions or organizations are some examples of words that you may choose to eliminate, depending on the purpose of your corpus and the needs of your students. The *Find and Replace* function can be helpful in making sure all instances of particular words are deleted, by replacing words to be eliminated with a space. After completing word elimination, and prior to analysis, corpus files must be converted to txt format, preferably in Unicode (UTF-8) text encoding.

Text Analysis Software: *AntConc*

Many software programs can be used for text analysis. A Canadian initiative, the *Compleat Lexical Tutor* website (Cobb, n.d.), offers a multitude of computer programs and services for learners, researchers, and teachers, including vocabulary profiling, word concordancers, and frequency calculators. It is a free resource that requires familiarization prior to understanding all of its uses. *Sketch Engine* (www.sketchengine.co.uk) is also recommended, but requires a monthly subscription. *AntConc* (Anthony, 2014) is the most comprehensive and easy-to-use freely available corpus analysis software for concordance and text analysis that I have found. The *AntConc* webpage (http://www.laurenceanthony.net/software/antconc/) includes links to video tutorials and discussion groups. *AntConc* is available for Windows, Macintosh, and Linux computers. For these reasons, it is good software for teachers developing a corpus for the first time. In the following section, how to use *AntConc* to develop a frequency-based vocabulary list will be explained. The screenshots provided are from the most recent version of *AntConc* for Macintosh for OS 10.x: 3.4.4m.

Preparing to Use AntConc

The *AntConc* software must first be downloaded and installed from the *AntConc* webpage. A file called *AntBNC Lemma List* must also be downloaded from the *Lemma List* section at the bottom of the page. Finally, the corpus txt files are needed.

Creating a Frequency List

A frequency list of *lemmas,* or headwords as found in a dictionary (McEnery & Hardie, 2012), can be generated by completing the following steps.

1. Launch *AntConc*.

2. Upload your txt corpus file(s): Go to *File*, and select *Open File(s)* from the dropdown menu (Figure 1). This brings you to another window where the file(s) can be selected from your computer. After this is done, click *Open* (Figure 2). The corpus file(s) will then show as loaded on the left under *Corpus Files* (Figure 4).

3. Set the token definition: Go to *Settings* and select *Global Settings* from the dropdown menu. Select the *Token Definition* category. The *Letter* box should automatically be checked under *Letter Token Classes*. Next, under

User-Defined Token Class, check *Append Following Definition*. Then type an apostrophe ('), followed by a hyphen (-). No space is needed between them. Finally, click *Apply* at the bottom (Figure 3). Please note that this step must be done prior to uploading the *AntBNC Lemma List* file.

4. Upload the *AntBNC Lemma List* file: The steps necessary to complete this process are shown in Figures 4 to 7. First, go to *Settings* and select *Tool Preferences* from the dropdown menu (Figure 4). In the *Tool Preferences*

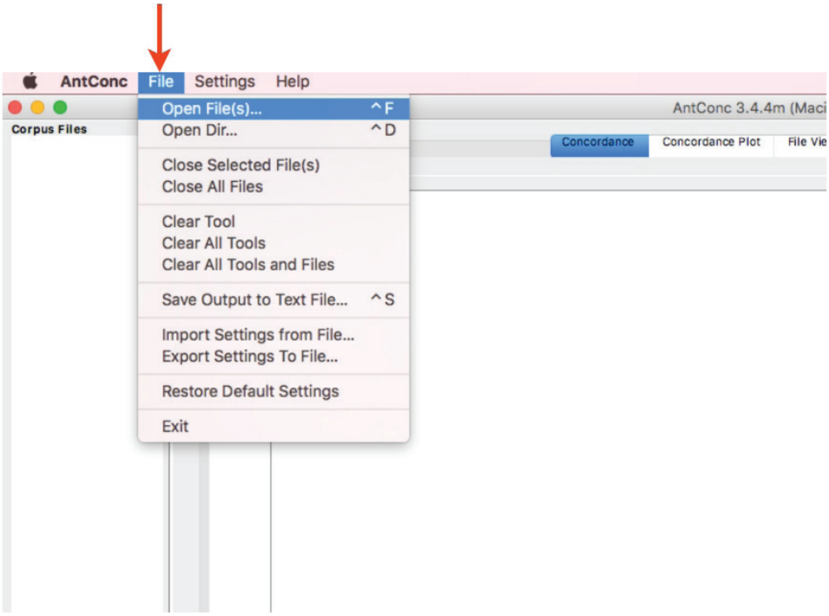


Figure 1: Uploading a corpus file

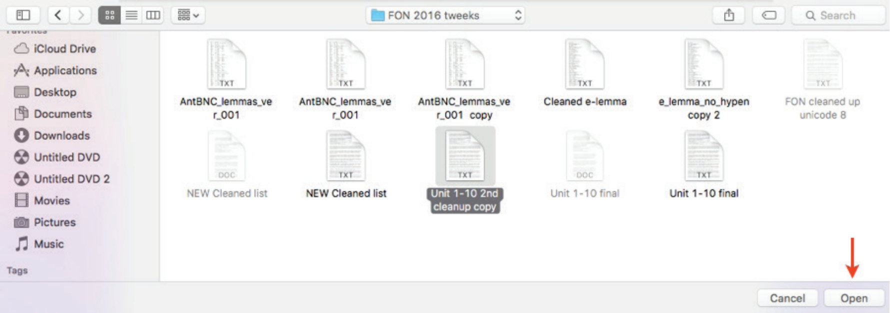
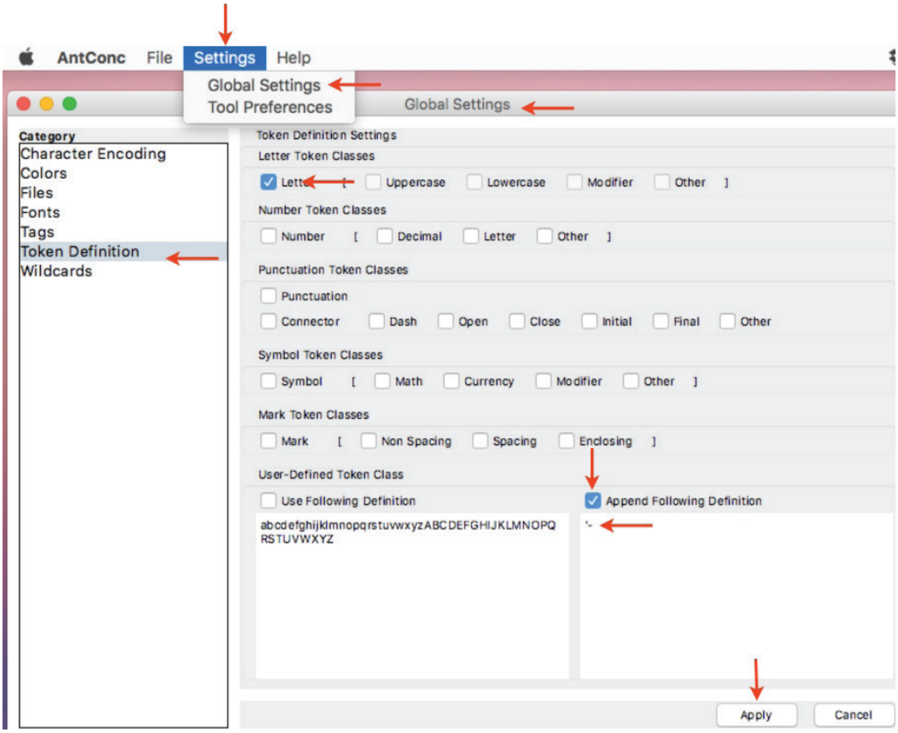


Figure 2: Selecting a file



Token Definition Settings

Letter Token Classes

☑ Letter [□ Uppercase □ Lowercase □ Modifier □ Other]

Number Token Classes

□ Number [□ Decimal □ Letter □ Other]

Punctuation Token Classes

□ Punctuation

□ Connector □ Dash □ Open □ Close □ Initial □ Final □ Other

Symbol Token Classes

□ Symbol [□ Math □ Currency □ Modifier □ Other]

Mark Token Classes

□ Mark [□ Non Spacing □ Spacing □ Enclosing]

User-Defined Token Class

□ Use Following Definition

abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSTUVWXYZ

☑ Append Following Definition

'-

Figure 3: Setting the token definition

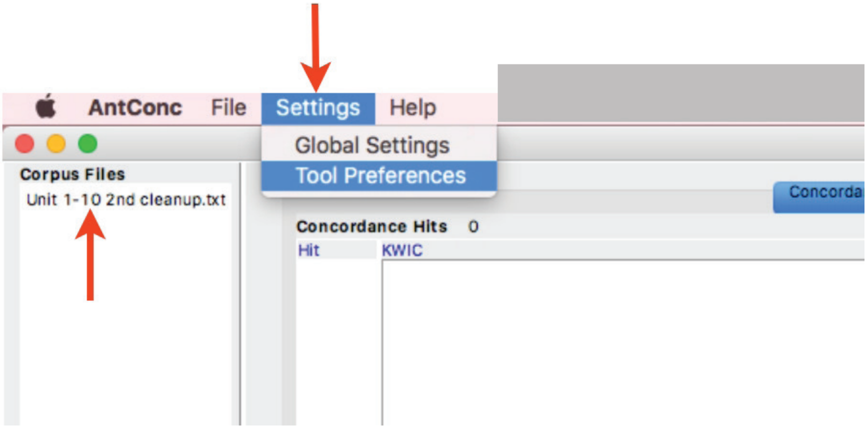
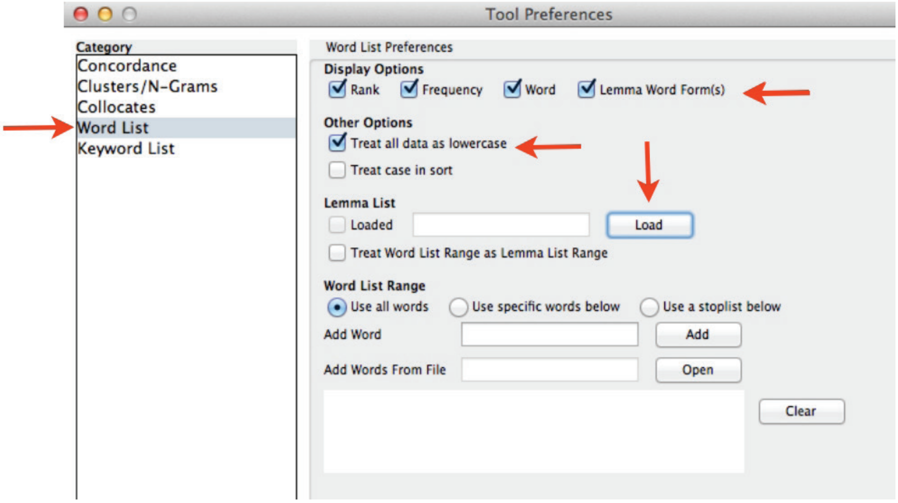


Figure 4: Uploading the e-lemma file (1)

window, select *Word List* under *Category* on the left. Use the default settings for *Display Options* (rank, frequency, word, lemma forms) and *Other Options* (treat all data as lowercase). Under *Lemma List* click on the *Load* button (Figure 5). This opens a window where the file can be selected. Click *Open* afterwards. Shortly after pressing the *Open* button, a *Lemma List Entries* window will appear (Figure 6). Click *OK*. After doing so, this window will disappear. The *Lemma List* will be indicated as *Loaded*. Finally, click *Apply* at the bottom (Figure 7).



Tool Preferences

Category
- Concordance
- Clusters/N-Grams
- Collocates
- Word List
- Keyword List

Word List Preferences

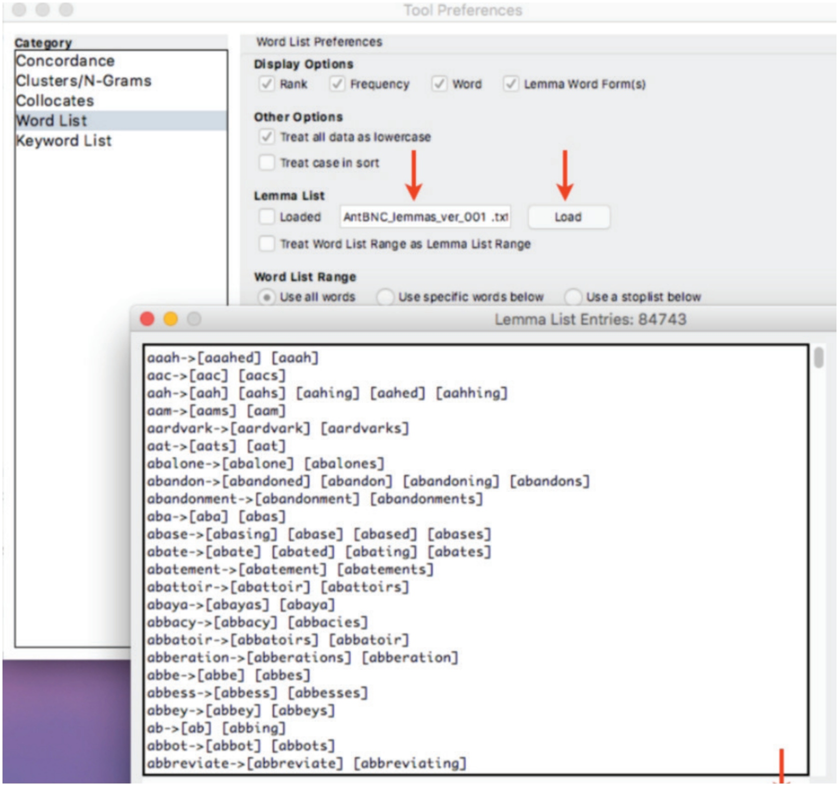
Display Options
☑ Rank ☑ Frequency ☑ Word ☑ Lemma Word Form(s)

Other Options
☑ Treat all data as lowercase
□ Treat case in sort

Lemma List
□ Loaded
□ Treat Word List Range as Lemma List Range

Word List Range
● Use all words ○ Use specific words below ○ Use a stoplist below
Add Word
Add Words From File

Figure 5: Uploading the e-lemma file (2)



Word List Preferences

Display Options

☑ Rank ☑ Frequency ☑ Word ☑ Lemma Word Form(s)

Other Options

☑ Treat all data as lowercase

☐ Treat case in sort

Lemma List

☐ Loaded AntBNC_lemmas_ver_001 .txt

☐ Treat Word List Range as Lemma List Range

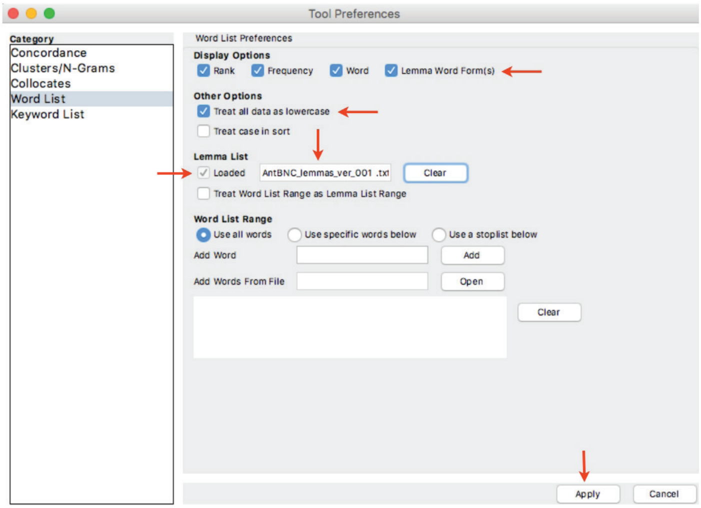
Word List Range

● Use all words ○ Use specific words below ○ Use a stoplist below

Lemma List Entries: 84743

```
aaah->[aaahed] [aaah]
aac->[aac] [aacs]
aah->[aah] [aahs] [aahing] [aahed] [aahhing]
aam->[aams] [aam]
aardvark->[aardvark] [aardvarks]
aat->[aats] [aat]
abalone->[abalone] [abalones]
abandon->[abandoned] [abandon] [abandoning] [abandons]
abandonment->[abandonment] [abandonments]
aba->[aba] [abas]
abase->[abasing] [abase] [abased] [abases]
abate->[abate] [abated] [abating] [abates]
abatement->[abatement] [abatements]
abattoir->[abattoir] [abattoirs]
abaya->[abayas] [abaya]
abbacy->[abbacy] [abbacies]
abbatoir->[abbatoirs] [abbatoir]
abberation->[abberations] [abberation]
abbe->[abbe] [abbes]
abbess->[abbess] [abbesses]
abbey->[abbey] [abbeys]
ab->[ab] [abbing]
abbot->[abbot] [abbots]
abbreviate->[abbreviate] [abbreviating]
```

Figure 6: Uploading the e-lemma file (3)



Word List Preferences

Display Options

☑ Rank ☑ Frequency ☑ Word ☑ Lemma Word Form(s)

Other Options

☑ Treat all data as lowercase

☐ Treat case in sort

Lemma List

☑ Loaded AntBNC_lemmas_ver_001 .txt

☐ Treat Word List Range as Lemma List Range

Word List Range

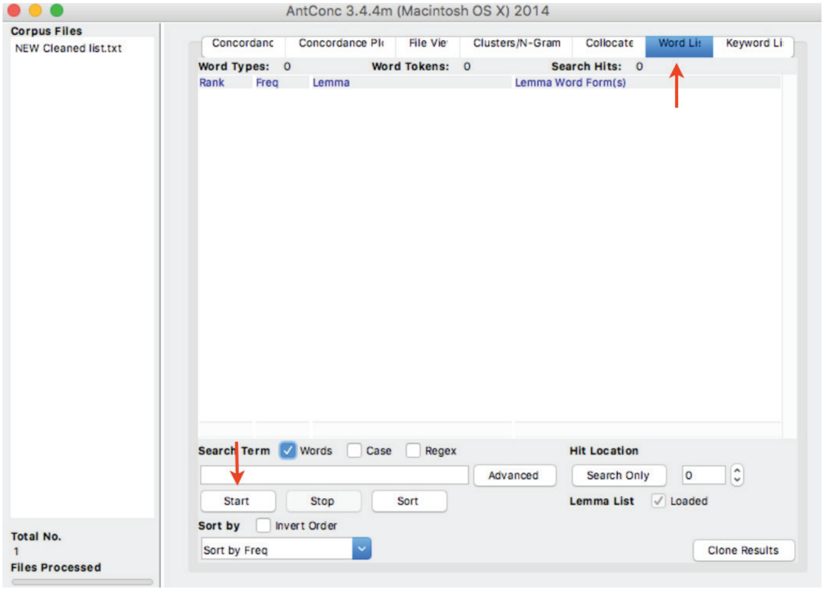
● Use all words ○ Use specific words below ○ Use a stoplist below

Add Word

Add Words From File

Figure 7: Uploading the e-lemma file (4)

5. Generate the frequency list: Select *Word List* at the top right of the navigation bar, and click *Start* (Figure 8). The frequency list will appear within a few seconds (Figure 9).



Word Types: 0 Word Tokens: 0 Search Hits: 0

☑ Words □ Case □ Regex

0

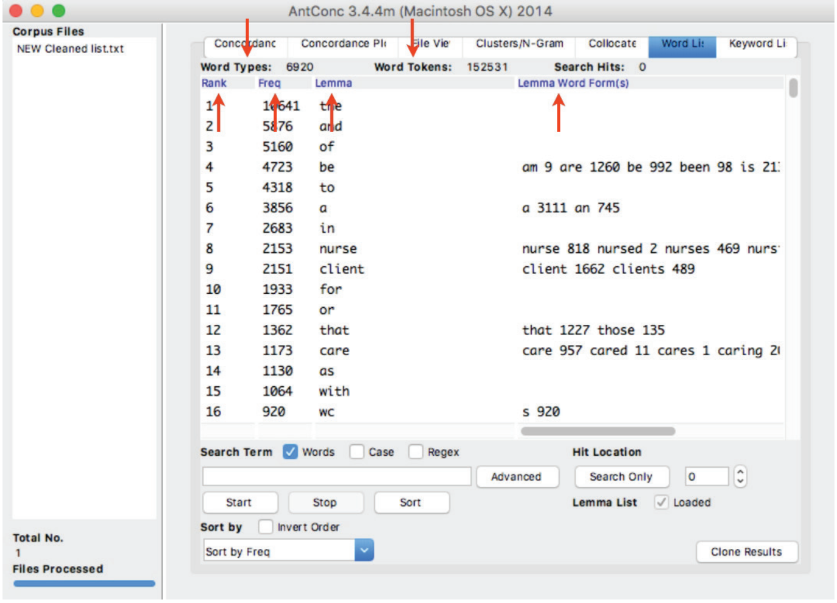
Lemma List ✓ Loaded

□ Invert Order

Sort by Freq

Total No. 1

Figure 8: Generating a frequency list



Word Types: 6920 Word Tokens: 152531 Search Hits: 0

Rank	Freq	Lemma	Lemma Word Form(s)
1	10641	the	
2	5876	and	
3	5160	of	
4	4723	be	am 9 are 1260 be 992 been 98 is 21
5	4318	to	
6	3856	a	a 3111 an 745
7	2683	in	
8	2153	nurse	nurse 818 nursed 2 nurses 469 nurs
9	2151	client	client 1662 clients 489
10	1933	for	
11	1765	or	
12	1362	that	that 1227 those 135
13	1173	care	care 957 cared 11 cares 1 caring 2
14	1130	as	
15	1064	with	
16	920	wc	s 920

☑ Words □ Case □ Regex

0

Lemma List ✓ Loaded

□ Invert Order

Sort by Freq

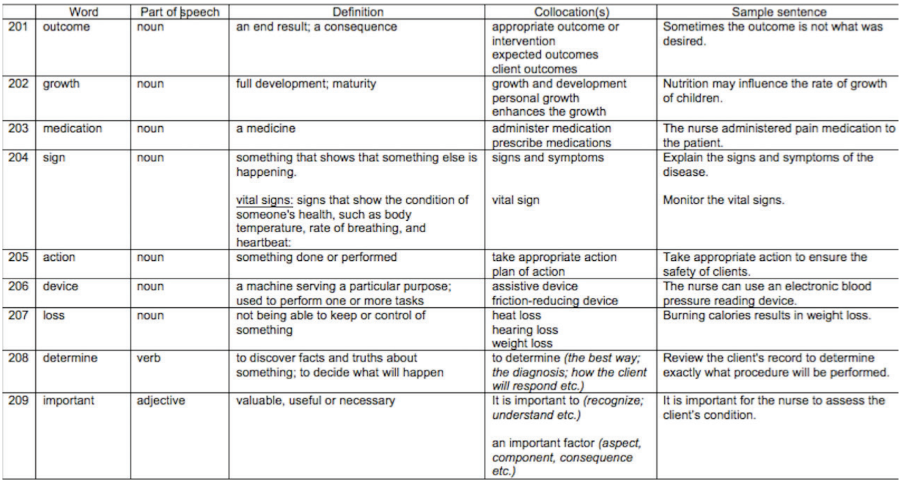
Total No. 1

Figure 9: Completed frequency list

Key data indicators generated by *AntConc* are shown in Figure 9. They first include the number of word types (number of base words, or lemmas) and word tokens (total word count) in the corpus. In the main data box, the data are presented in four columns, including rank, frequency (how many times the word was found), lemma (or word as in a dictionary entry, such as *eat*), and corresponding lemma forms (various inflections of a lemma that do not change the meaning, such as *eats*, *eating*, *ate*). For each lemma form, the number of instances found in the corpus is given in parentheses. It is possible to copy and paste the data in each column, one column at a time, into an Excel spreadsheet if needed.

Developing an Annotated Frequency-based Vocabulary List

An annotated vocabulary list, including frequency, part of speech, definition, collocation, and sample sentence, can be an important tool for students. This allows students to study words in order of frequency, with a focus on how the words are predominantly used in the target corpus, thus making the learning process more efficient. Figure 10 shows part of an annotated vocabulary list developed from our corpus.



	Word	Part of speech	Definition	Collocation(s)	Sample sentence
201	outcome	noun	an end result; a consequence	appropriate outcome or intervention expected outcomes client outcomes	Sometimes the outcome is not what was desired.
202	growth	noun	full development; maturity	growth and development personal growth enhances the growth	Nutrition may influence the rate of growth of children.
203	medication	noun	a medicine	administer medication prescribe medications	The nurse administered pain medication to the patient.
204	sign	noun	something that shows that something else is happening. vital signs: signs that show the condition of someone's health, such as body temperature, rate of breathing, and heartbeat:	signs and symptoms vital sign	Explain the signs and symptoms of the disease. Monitor the vital signs.
205	action	noun	something done or performed	take appropriate action plan of action	Take appropriate action to ensure the safety of clients.
206	device	noun	a machine serving a particular purpose; used to perform one or more tasks	assistive device friction-reducing device	The nurse can use an electronic blood pressure reading device.
207	loss	noun	not being able to keep or control of something	heat loss hearing loss weight loss	Burning calories results in weight loss.
208	determine	verb	to discover facts and truths about something; to decide what will happen	to determine *(the best way; the diagnosis; how the client will respond etc.)*	Review the client's record to determine exactly what procedure will be performed.
209	important	adjective	valuable, useful or necessary	It is important to *(recognize; understand etc.)* an important factor *(aspect, component, consequence etc.)*	It is important for the nurse to assess the client's condition.

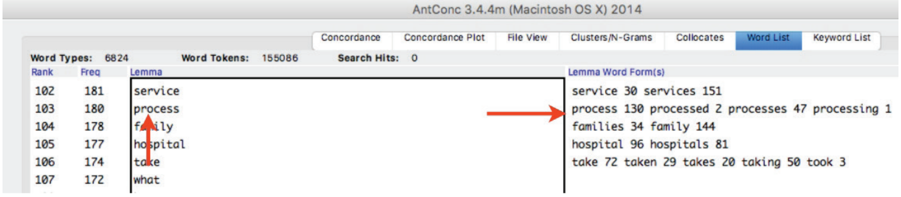
Figure 10: Annotated word list example

Part of Speech

In developing our annotated list, it was decided to only indicate the prominent part of speech according to how the words were used in the corpus. Two parts of speech were noted when the word was used as such. *AntConc* does not automatically identify word forms. Other paid concordance software programs, such as *Sketch Engine*, do. To recognize the prominent part of speech of a lemma word using *AntConc*, first look at the *Lemma Word Form(s)* column (see Figure 11). For example, in our corpus the lemma *process* included the following lemma forms: *process* (130), *processed* (2), *processes* (47), *processing* (1).

The words *process* and *processes* could both be used as nouns or verbs in the corpus. In such a case, you must then look at all the sentences where the word is used. Clicking on a word in the *Lemma* column will allow you to see all the sentences where it is found in the corpus (Figure 12). These are called *concordance lines*.

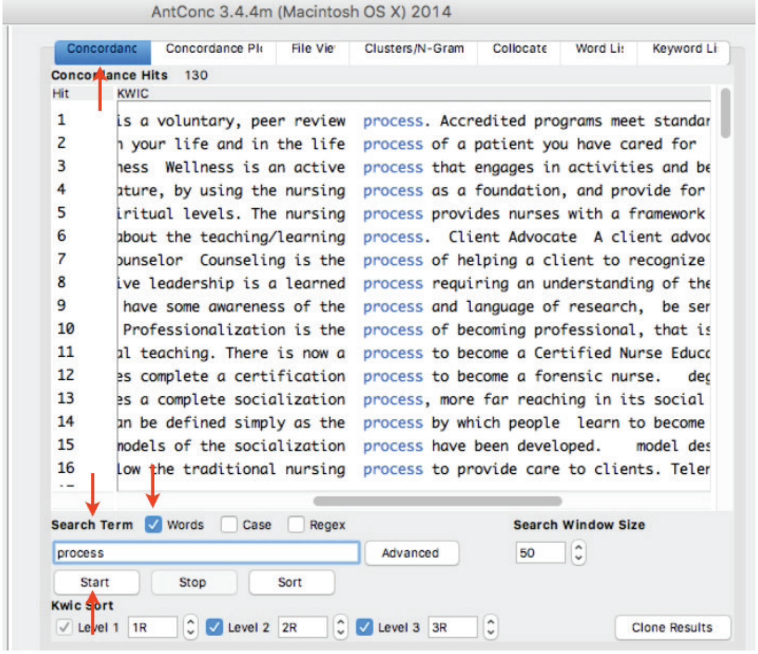
To see the sentences where one of the lemma word forms is found, type the word in the *Search Term* box (with the *Words* box selected), and press *Start* (Figure 12). Repeat this step for each of the lemma word forms.



Word Types: 6824 Word Tokens: 155086 Search Hits: 0

Rank	Freq	Lemma	Lemma Word Form(s)
102	181	service	service 30 services 151
103	180	process	process 130 processed 2 processes 47 processing 1
104	178	family	families 34 family 144
105	177	hospital	hospital 96 hospitals 81
106	174	take	take 72 taken 29 takes 20 taking 50 took 3
107	172	what	

Figure 11: Lemma and lemma forms



Concordance Hits 130

```
Hit  KWIC
1    is a voluntary, peer review   process. Accredited programs meet standar
2    n your life and in the life   process of a patient you have cared for
3    ness  Wellness is an active   process that engages in activities and be
4    ature, by using the nursing   process as a foundation, and provide for
5    iritual levels. The nursing   process provides nurses with a framework
6    about the teaching/learning   process.  Client Advocate  A client advoc
7    ounselor  Counseling is the   process of helping a client to recognize
8    ive leadership is a learned   process requiring an understanding of the
9     have some awareness of the   process and language of research,  be ser
10    Professionalization is the   process of becoming professional, that is
11   al teaching. There is now a   process to become a Certified Nurse Educa
12   es complete a certification   process to become a forensic nurse.   deg
13   es a complete socialization   process, more far reaching in its social
14   an be defined simply as the   process by which people  learn to become
15   models of the socialization   process have been developed.     model des
16   Low the traditional nursing   process to provide care to clients. Teler
```

Search Term ☑ Words □ Case □ Regex

process

Advanced

Search Window Size 50

Start Stop Sort

Kwic Sort

☑ Level 1 1R ☑ Level 2 2R ☑ Level 3 3R

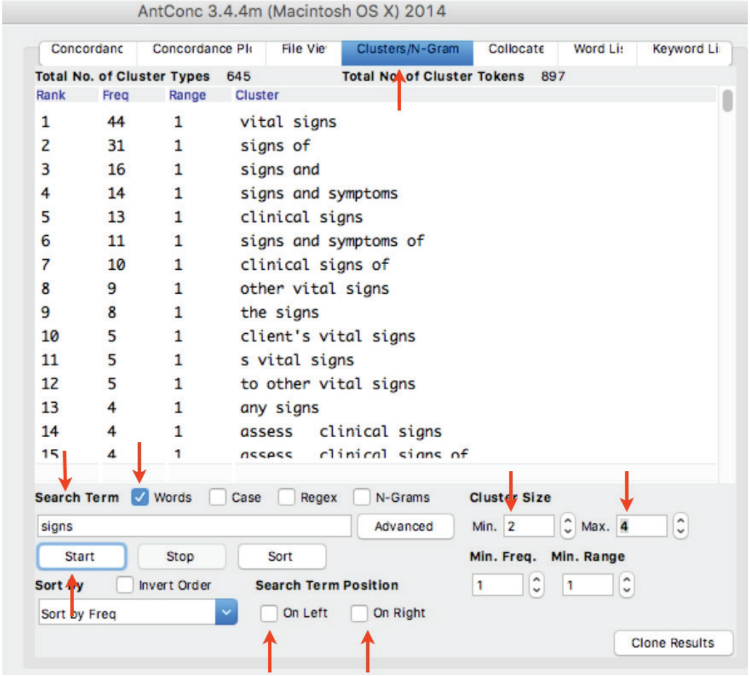
Clone Results

Figure 12: Corpus concordance lines

By looking at all the concordance lines, we found that the words *process* and *processes* were mostly used as nouns in our corpus. It was decided, however, to include both noun and verb as parts of speech, as the verb forms were used more than 50 times.

Definition

Using definitions from a unified source helps students because the definitions tend to follow the same pattern of presentation. *The Cambridge Learner's Dictionary Online* (http://dictionary.cambridge.org/dictionary/learner-english/) is a helpful tool as the definitions are written for language learners. In keeping with our goal of focusing on the contextual meanings and uses of words, attention must be given to extracting the salient sense of each word as used in the corpus. This was done when looking at the part of speech and concordance lines in the step explained above. Corresponding definitions were then taken from *The Cambridge Learner's Dictionary Online*. When a high-frequency collocation had a meaning of its own, the definition was also included. For example, the highest frequency collocation for the word "sign" in our corpus was "vital signs" (Figure 13). *The Free Medical Dictionary* (http://medical-dictionary.thefreedictionary.com/), also web-based, was used for definitions of more technical language not covered in the *The Cambridge Learner's Dictionary Online*.



Total No. of Cluster Types 645
Total No. of Cluster Tokens 897

Rank	Freq	Range	Cluster
1	44	1	vital signs
2	31	1	signs of
3	16	1	signs and
4	14	1	signs and symptoms
5	13	1	clinical signs
6	11	1	signs and symptoms of
7	10	1	clinical signs of
8	9	1	other vital signs
9	8	1	the signs
10	5	1	client's vital signs
11	5	1	s vital signs
12	5	1	to other vital signs
13	4	1	any signs
14	4	1	assess clinical signs
15	4	1	assess clinical signs of

Search Term ☑ Words □ Case □ Regex □ N-Grams
signs
Cluster Size Min. 2 Max. 4
Min. Freq. 1 Min. Range 1
Sort by: Sort by Freq □ Invert Order
Search Term Position □ On Left □ On Right

Figure 13: Collocations

Collocations

To understand how a word is used in the specific context of the target corpus, all words and their associated lemma form(s) should be checked for collocations. The following steps will help you to identify high-frequency collocations for inclusion in the annotation (Figure 13). First, select *Cluster/N-Grams* from the top navigation bar. With the *Words* box selected, type a word in the *Search Term* box. Adjust the *Cluster Size*: *Min. 2* and *Max. 4* were used in our project, as collocations of 5 words or more are often not the highest in frequency. Finally, you can select a *Search Term Position* (*On Left* or *On Right*). By not selecting a *Search Term Position,* collocations, including words both to the left and right, will be shown. After selecting the above settings, click *Start*. Repeat the process for each lemma form.

Sample Sentences

The *Part of Speech* section above included an explanation of how to access sample sentences. Choose a sentence that will be easy to understand. Sample sentences may need to be simplified and shortened for students' ease of understanding. In our project, an effort was also made to use corpus words previously seen at higher frequencies (lower rank number) in sample sentences to create repetition. This gives the learners a chance to review previous vocabulary, and it helps to make the meaning of the target words clear.

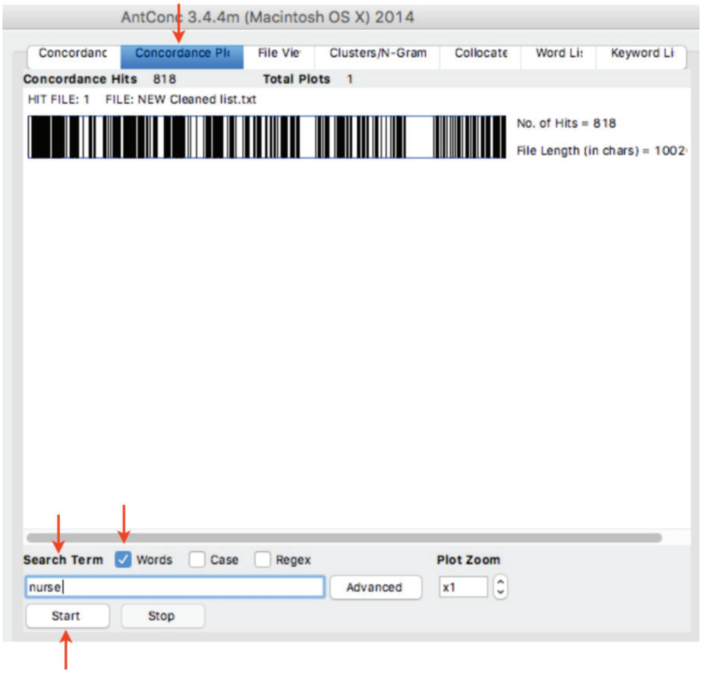
Other Useful *AntConc* Functions

Calculating Coverage

As noted previously, lexical coverage is important for reading comprehension. While *AntConc* does not directly calculate coverage, you can do so in a few steps. For instance, you might be interested in calculating the coverage for the first 1,000 words of your corpus. To do this, first generate a word list (Figure 9), and go to the *Freq*. column. Then, copy and paste the frequencies listed for the first 1,000 words into an Excel document. Use Excel to calculate the sum of all these frequencies. Finally, divide this sum by the number of word tokens in your corpus.

Concordance Plot

The concordance plot function shows a physical representation of the dispersion of a target word in the corpus. This can help to identify words that may only appear in a section of a corpus due to the specific topic matter. Even distribution of a word across the corpus indicates that the word is less likely to be topic- or chapter-specific. To access a concordance plot, you can first click a word in the Lemma column under *Word List* in the top navigation bar, and then click on the *Concordance Plot* in the top navigation bar. You can also type a word directly into the *Search Term* box (with the *Words* box selected) under *Concordance Plot* and press *Start* (Figure 14).



AntConc 3.4.4m (Macintosh OS X) 2014

Concordance Hits 818 Total Plots 1

HIT FILE: 1 FILE: NEW Cleaned list.txt

No. of Hits = 818
File Length (in chars) = 1002

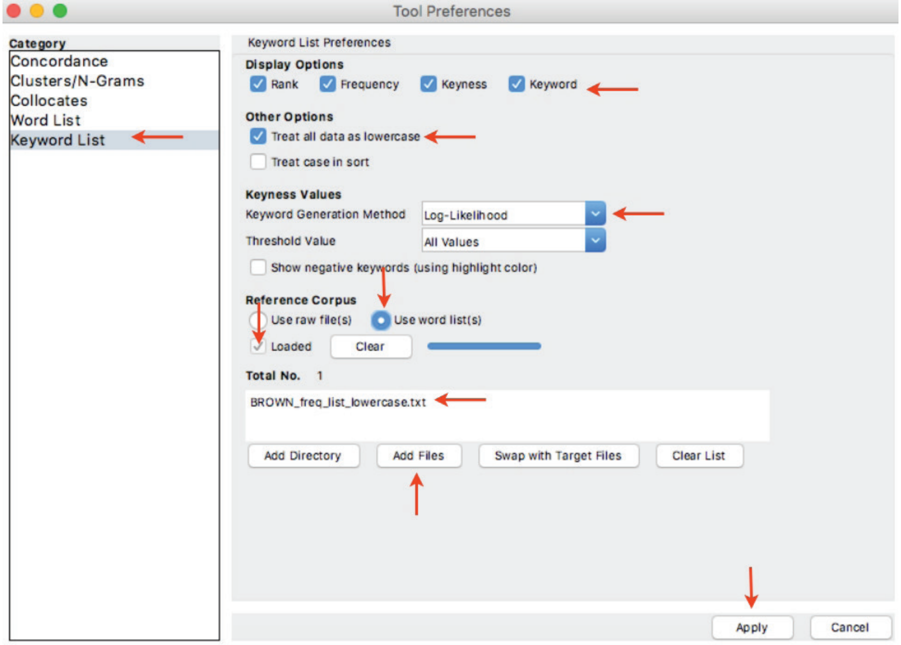
Search Term ☑ Words □ Case □ Regex

nurse

Figure 14: Concordance plot

Keyword Lists

A *keyword* refers to a word that occurs more (or less) frequently in a given corpus compared to a reference corpus of general English (McEnery et al., 2006). Overrepresentation is often an indication that the word is specific to the field of study, but may also indicate a bias due to the topic matter (Millar & Budgell, 2008). Your corpus can be compared to, for example, the *British National Corpus* (written, spoken, and combined) or the *Brown Corpus* (Francis & Kucera, 1964). These lists can be downloaded from the *AntConc* webpage under *Word Frequency Lists* at the bottom of the page. To do this analysis, you must first create a frequency list. Next, go to *Settings* and *Tool Preferences* (as in Figure 4). From there, choose the *Keyword List* under *Category* on the left. Basic settings are shown in Figure 15. You must choose the text file by clicking *Add Files* (found toward the bottom of the page), then click *Load*. Once the file is loaded, a check will appear in the *Loaded* box. Finally, click *Apply*. Then, go back to the *Keyword List* section on the top navigation bar, and press *Start*. The resulting list shows the words that are of unusually high frequency in your target corpus. The word “nurse” and “client” were the two highest frequency keywords in our corpus, showing the overrepresentation, and hence importance, of those words in it (Figure 16).



Display Options

☑ Rank ☑ Frequency ☑ Keyness ☑ Keyword

Other Options

☑ Treat all data as lowercase

□ Treat case in sort

Keyness Values

Keyword Generation Method: Log-Likelihood

Threshold Value: All Values

□ Show negative keywords (using highlight color)

Reference Corpus

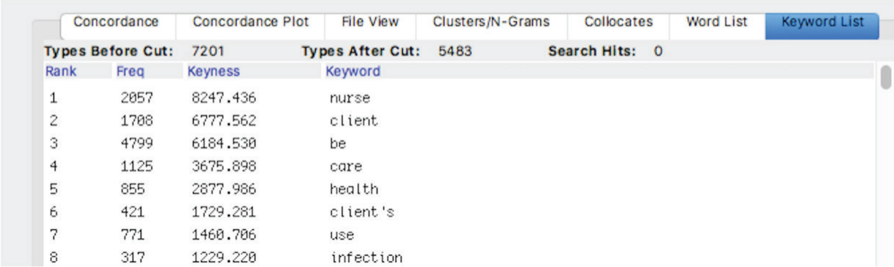
○ Use raw file(s) ● Use word list(s)

☑ Loaded

Total No. 1

BROWN_freq_list_lowercase.txt

Figure 15: Keyword function settings



Types Before Cut: 7201 Types After Cut: 5483 Search Hits: 0

Rank	Freq	Keyness	Keyword
1	2057	8247.436	nurse
2	1708	6777.562	client
3	4799	6184.530	be
4	1125	3675.898	care
5	855	2877.986	health
6	421	1729.281	client's
7	771	1460.706	use
8	317	1229.220	infection

Figure 16: List of keywords

Application

The individualized vocabulary acquisition program established in our EAP program follows an interval learning approach (also called spaced repetition) based on Ebbinghaus’s (1885/1964) learning and forgetting curve. With respect to explicit learning, memory research has consistently found that learning using spaced repetition leads to better long-term retention than learning via massed presentation (Nation, 2013).

Electronic flashcards (Shimoda et al., 2016) were first used. Students, however, voiced a preference for handmade paper flashcards. Handmade flashcards and a spaced repetition technique similar to the one described by Mondria and Mondria-De Vries (1994) are now being used by most students. Our approach also incorporates many of Nation's (2013) recommendations regarding making and using word cards (pp. 445–454).

To address individual language gaps, students are asked to go through the lists in order of frequency and self-select words to study. This is done at a rate of 10 to 15 new words per week. One card is made for each word. The target word is written on one side, and the remaining information from the annotated list is written on the other. Students are tested orally one-on-one, weekly or biweekly. Teachers randomly select approximately five flashcards. For each word, students are asked to provide the meaning, part of speech, and collocation and/or sample sentence as given in the annotated lists. A one-on-one approach is a great way to clarify the meaning of words. These tests account for 5–10% of students' *Academic Reading* course grade. As students invest a considerable amount of time and effort making the cards and learning the words, grading is done leniently. Testing is cumulative within a level, and across EAP levels. In other words, students keep reviewing the words over the course of their study time in the EAP program. This helps to ensure long-term retention of the meanings and uses of words. A more detailed description of spaced repetition and ways to use flashcards for vocabulary learning will be the topic of a future article. For more ideas on how to use word lists in the classroom, see Nation (2016).

Closing Comments

Embracing a corpus-based approach to vocabulary teaching and learning can be a fulfilling and rewarding experience for teachers and students alike. Using an annotated frequency-based vocabulary list made from a specialized corpus can help maximize student learning for study time spent by focusing on the most useful words to study and on how these words are used in that specific field or context. These lists can serve as a basis not only for creating a vocabulary syllabus, but also for creating contextualized materials and developing other language learning activities. This article was written to help teachers with minimal corpus linguistics experience become able to develop a specialized corpus and annotated frequency-based vocabulary lists for classroom use. It is hoped that the explanations provided will serve to open the door to the field of corpus linguistics and inspire teachers to attempt such a task.

Acknowledgements

I would like to acknowledge and thank Jody Shimoda and William Kay for giving me the opportunity to work on their team for the nursing corpus project. Many thanks also go to Dr. Charles Browne for his support and encouragement, and Dr. Laurence Anthony for his technical support. Finally, I sincerely thank the anonymous reviewers for their valuable feedback.

The Author

Marie-Claude Toriida, MAT, is an EAP Instructor in the Nursing Foundation Program at the University of Calgary in Qatar. Her research interests include corpus development, corpus-based vocabulary teaching, vocabulary learning strategies, and reading comprehension.

References

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*(4), 425–438. https://doi.org/10.1093/applin/24.4.425

Adolphs, S. & Schmitt, N. (2004). Vocabulary coverage according to spoken context. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 39–52). Amsterdam, Netherlands: John Benjamins.

Anthony, L. (2014). *AntConc* (Version 3.3.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software/antconc/

Anthony, L. (2015). *AntfileConverter* [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software/antfileconverter/

Bin Baki, R., & Kameli, S. (2013). The impact of vocabulary knowledge level on EFL reading comprehension. *International Journal of Applied Linguistics and Literature*, *2*(1), 85–89. https://doi.org/10.7575/ijalel.v.2n.1p85

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, *36*, 1–22. https://doi.org/10.1093/applin/amt018

British National Corpus, version 3 (BNC XML Edition). (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available from http://www.natcorp.ox.ac.uk

Browne, C., Culligan, B., & Phillips, J. (2013a). The new academic world list. Available from http://www.newgeneralservicelist.org/nawl-new-academic-word-list/

Browne, C., Culligan, B., & Phillips, J. (2013b). The new general service list. Available from http://www.newgeneralservicelist.org/

Budgell, B., Miyazaki, M., O'Brien, M., Perkins, R., & Tanaka, Y. (2007). Developing a corpus of the nursing literature: A pilot study. *Japan Journal of Nursing Science*, *4*(1), 21–25. https://doi.org/10.1111/j.1742-7924.2007.00071.x

Cobb, T. (n.d.). *The compleat lexical tutor*. Retrieved from http://www.lextutor.ca/

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Coxhead, A. (2011). The academic word list 10 years on: Research and teaching implications. *TESOL Quarterly*, *45*(2), 355–362. https://doi.org/10.5054/tq.2011.254528

Davies, M. (2008-) *The Corpus of Contemporary American English: 520 million words, 1990–present*. Available from http://corpus.byu.edu/coca/

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology.* New York, NY: Dover. (Original work published 1885)

Francis, W. N., & Kucera, H. (1964). *Brown corpus*. Providence, RI: Department of Linguistics, Brown University.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*, 305–327. https://doi.org/10.1093/applin/amt015

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly, 41*(2), 235–253. https://doi.org/10.1002/j.1545-7249.2007.tb00058.x

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–30.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Clevedon, UK: Multilingual Matters.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1), 15–30.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics.* Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

McEnery, T., Xiao, Z., & Tono, Y. (2006). *Corpus based language studies: An advanced resource book*. London, UK: Routledge. https://doi.org/10.1017/s0047404508080615

Millar, N., & Budgell, B. N. (2008). The language of public health: A corpus based analysis. *Journal of Public Health*, *16*(5), 369–374. https://doi.org/10.1007/s10389-008-0178-9

Mondria, J. A., & Mondria-De Vries, S. (1994). Efficiently memorizing words with the help of word cards and "hand computer": Theory and applications. *System*, *22*(1), 47–57. https://doi.org/10.1016/0346-251X(94)90039-6

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge: Cambridge University Press.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing.* Amsterdam, Netherlands: John Benjamins. https://doi.org/10.1075/z.208

Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (pp. 6–19). Cambridge, UK: Cambridge University Press.

Nelson, M. (2010). Building a written corpus: What are the basics? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 53–65). London, UK: Routledge.

Neufeld, S., & Billuroğlu, A. (2005). *In search of the critical lexical mass: How 'general' is the GSL? How 'academic' is the AWL?* Retrieved from https://www.academia.edu/573862/In_search_of_the_critical_lexical_mass_How_general_is_the_GSL_How_academic_is_the_AWL

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal*, *95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Shimoda, J., Toriida, M.-C., & Kay, D. W. (2016). Improving learning outcomes: Creating and implementing a specialized corpus. *Perspectives*, *24*(1), 22–28.

West, M. (1953). *A general service list of English words*. London, UK: Longman.

Yang, M. N. (2015). A nursing academic word list. *English for Specific Purposes*, *37*, 27–38. https://doi.org/10.1016/j.esp.2014.05.003