**PASAA**
**Volume 52**
**July - December 2016**

# Development and Use of a Corpus Tailored for Legal English Learning

**Jason Skier**
**Jutarat Vibulphol**
*Chulalongkorn University*

**Abstract**

While corpus linguistics has been applied towards many specific academic purposes, reports are few regarding its use to facilitate learning of legal English by non-native English speakers. Specialized corpora are required because legal English often differs significantly from ordinary usage, with words such as bar, motion, and hearing having completely different meanings and use. This paper documents the process of creating and validating a sixteen million-word corpus of (American) legal English, and provides examples of analyses available for language learners. Written decisions and oral argument transcripts from the U.S. Supreme Court and other appellate courts were ultimately chosen to comprise the corpus due to their authentic and comprehensive use of legal jargon. Overall, this corpus demonstrates that appellate court decisions, available online, can comprise a corpus tailored for legal English learning.

**Keywords**: Corpus, legal English, English as a foreign language, AntConc

**Development and Use of a Corpus Tailored for Legal English Learning**

When most lawyers hear the word "corpus", they typically think of the *habeas corpus* proceeding, which demands that the body of a prisoner be brought before a judge; while *corpus linguistics* examines a body of text to research relationships between words. Legal English frequently differs from general English in both meaning and usage. For example, the word 'hearing' typically refers to a person's ability to perceive sound. In legal English, however, the same word describes a court session where attorneys may present evidence and legal arguments to persuade a tribunal. Accordingly, the phrase 'I lost my hearing yesterday' conveys completely a different meaning in legal English versus general English. Therefore, attempting to use non-legal corpora to research legal English usage will not yield meaningful results.

Using the legal English corpus, anyone looking to accurately describe judicial error could use collocation to reveal that "manifest", "plain", "reversible" and "harmless" are frequently discussed with "error". Investigation will further reveal that these are all discrete types of errors, and that the correct nomenclature must be used to convey the proper legal meaning (Stith, 1990). Correct use of legal English nomenclature is vitally important for all attorneys. Using incorrect terms fails to give adversaries proper notice, and failing to give proper notice may constitute waiver of an issue (Hawaii Rules of Appellate Procedure 28(b)(7), 1984). This requirement is strictly followed because, "courts are not gambling halls, but forums for the discovery of truth" (State v. Haanio, 2001). Accordingly, creating a corpus specifically for legal English would be useful for learning proper use of English in the legal context, especially for non-native English speakers.

While the benefits of using corpora in other types of language classrooms have been well documented (e.g. Crosthwaite, 2012; O'Keeffe, McCarthy, & Carter, 2007; Yoon & Hirvela, 2004), the use of corpora to assist legal English learning has not been widely explored (Fan & Xunfeng, 2002; Hafner & Candlin, 2007; Hunston, 2002). For

this project, a corpus was created using published court decisions regularly used by attorneys in the United States to study legal issues. This paper also shows how language learners can analyze this corpus using concordance, collocation, and chunking to discover legal English lexicon.

## Literature Review

### *The lexical approach*

This work was motivated by the principles of the lexical approach proposed by Lewis (1993), which considers language to be more than random combinations of single words but rather non-random combinations of specific phrases and idioms, called lexicon. Lexicon consists of frequently produced chunks of language combining to produce coherent communication (Lewis, 1993). Central to this idea is collocation, or the examination of which words frequently appear in close proximity to other words. These repeating collocated chunks, such as 'burden of proof' or 'cry wolf', are commonly found in authentic language use but not directly understood by examining individual words. The ability to comprehend and produce these lexical chunks is essential for language learning.

Corpus analysis is used to facilitate the learning of lexicon by examining authentic texts to show snippets of actual use of a search term, locating words commonly found in close proximity, and showing set phrases. Lewis (1993) suggests authentic texts because they contain natural jargon, word relationships, and fixed phrases commonly used in communication rather than artificial creations. This information produced by corpus analysis is useful for language learners because often these word relationships are not intuitive for non-native speakers. Computers are sometimes used to assist the analysis of bulk collections of texts to reveal these relationships.

### *Corpus Linguistics*

Corpus linguistics leverages the power of computers to analyze large collections of authentic texts to locate language chunks, and was

therefore recommended by Lewis (1997) as a tool to facilitate the implementation of the lexical approach. Because analyzing large volumes of text requires significant computer processing power, the first significant attempt at using corpus linguistics for foreign language learning was not until the 1987 publication of the Collins Birmingham University International Language Database, commonly referred to as COBUILD. Continually updated to the present day, the full COBUILD corpus currently contains 4.5 billion words (Collins, 2013). The Bank of English, a subset containing 650 million words, is a widely used reference tool considered to be a representative sample of current, general English usage (Collins, 2013). The first major attempt at a corpus specifically for academic use was the Academic Word List, a 3.5 million-word corpus consisting of journals and textbooks from various native English speaking countries covering assorted academic subjects (Coxhead, 1998).

***User created corpora***

In addition to premade corpora described in the previous section, Lewis (2001) suggested that users might alternatively create corpora tailored for their specific academic purposes. However, when constructing corpora, users should be aware that including high quality materials in corpora yields better results than texts merely meeting the minimum criteria (Hyland, 1999; Reppen, 2004).

Few prior studies have discussed using corpora to facilitate learning legal English. Fan and Xunfeng (2002) documented the use of a bilingual corpus of Chinese and English law to assist translation by Hong Kong students. Hafner and Candlin (2007) gave legal writing students access to a corpus designed specifically for their course, via an online concordance tool designed to track their use.

The quality of any corpus created for specific academic purposes has traditionally been measured along two dimensions (Lewis, 1997). The first dimension measures whether source documents are proper authentic texts featuring relevant jargon. The second dimension measures whether the corpus contains enough documents to produce

meaningful results for searches. This second requirement ensures that the corpus contains plentiful and diverse examples of usage patterns containing the target jargon so that common phrases, modifiers, and action words will be revealed. However, corpora generated for specific academic purposes need not be as massive as the behemoths like COBUILD. A corpus for a specific purpose may contain less than 20 high quality documents and still yield representative examples of key words and collocations (Lewis, 2001). Additionally, Lewis also recommends that source materials not be edited in any way.

Corpus usefulness for language learning may also be measured for three purposes (Keck & Kim, 2014). First, "behind the scenes" refers to lesson and activity planning using the corpus as a reference, with student interaction with the corpus being optional. Using the corpus, research can be done prior to class to discover phrases and collocations common to the vocabulary being taught. This information can then be used when planning in-class activities. The "behind the scenes" method is also useful for situations where students may not yet have the computer skills required. Next, the corpus may be used as an in-class demonstration tool of language use and word relationships. This can be accomplished either by on-screen demonstrations, or having students independently run the software on their personal computers. Lack of requisite computer skills by either teachers or learners is always a concern when considering in-class use. Finally, the corpus may be used as a self-study tool for language learning.

## Development of the corpus

### *Design*

Based upon the recommendations of prior studies regarding user made corpora for specific academic purposes (e.g. Charles, 2014; Hafner & Candlin, 2007), this project aimed to construct a corpus of legal English to facilitate the learning of non-native speakers. To address the two dimensions of corpus quality suggested by Lewis (1997), the documents used to develop the corpus need to contain authentic legal English jargon, and be in sufficient quantity to produce

meaningful results. Corpora creation using actual attorney-client communication was initially considered. Unfortunately, compiling this data would certainly breach attorney-client confidentiality rules. The actual documents, including names and other confidential personal information, would be easily accessible for viewing in the corpus. Redaction would be an overwhelming task and would violate the prohibition against editing corpus materials (Lewis, 1997). Therefore, authentic client communications were not included in the corpus. Alternatively, U.S. Supreme Court published opinions are an ideal choice to comprise a corpus of legal English. Because all legal issues may come before the court, every area of law and associated jargon should be represented in their published judgments. Due to *stare decisis*, the entire country must follow the decisions of the United States Supreme Court. *Stare decisis* translates from Latin as "to stand by things decided." Following that rule, courts follow their own prior decisions on identical issues from other cases, as well as following prior decisions from superior tribunals. Because every word invokes *stare decisis*, a U.S. Supreme Court transcript or written judgment contains no superfluous verbiage, only succinct legal argument and analysis.

Additionally, written judgments from the highest state courts provide similar quality, and expand the corpus to include additional local differences in jargon. Appellate courts decisions from Hawaii, New York, and Florida were chosen due to their local differences, and the availability of appellate court decisions. Not only does this inclusion add massive amounts of high-quality authentic texts to the corpus, each jurisdiction has different lexicon for specific legal English terminology. Like the U.S. Supreme Court, these decisions are high quality because they represent the highest law in the jurisdiction and invoke *stare decisis*. Thus, state appellate legal decisions are also ideal, high quality texts to comprise a corpus of legal English.

***Creation***

First, a freeware program named AntConc was chosen to assist with the analysis of the created corpus. AntConc, available at http://www.laurenceanthony.net/software/antconc/ is available for Windows, OSX, and Linux (Anthony, 2014). The AntConc program was selected because it contains a full feature set, and is freely available. Whereas many other corpus analysis programs contain only a concordancing tool, AntConc also contains additional useful tools for examination, such as collocation and chunking. Prior studies involving academic use of corpora have also used AntConc (Charles, 2014; Csomay & Petrović, 2012).

After AntConc had been selected, the creation process next involved locating the source documents online, downloading, then converting into the proper format required by AntConc. Unfortunately, the format used by the courts to publish their decisions online is incompatible with AntConc. However, the creator of AntConc also provides AntFileConverter for bulk document conversion (Anthony, 2016).

After the U.S. Supreme Court, Florida, New York and Hawaii corpora were initially constructed in October 2015, the corpus contained over fourteen million word-tokens. As the corpus grew larger many search queries using AntConc, especially chunking, began to take significantly longer for the computer to process. Additionally, meaningful results were already being obtained for a full spectrum of civil and criminal law queries. Therefore, the decision was made to halt adding documents to the corpus. The completed corpus contains 4,697 documents with more than 16 million word tokens as shown in Appendix A. To enhance selection options and speed processing, the documents were divided into sub-corpora differentiating between jurisdictions, and transcripts or written judgments. Because AntConc software is able to simultaneously open multiple corpora, this allows the user to balance speed with comprehensiveness. Due to the smaller number of judgments, Florida and New York were combined into one

sub-corpus. See Appendix A for a detailed breakdown of the documents comprising the corpus.

***Validation***

Finally, three experts, instructors at major universities in Bangkok with Ph.D. degrees in their relevant fields, conducted the evaluation of the corpus created for this project. One expert currently teaches English for Law, another expert teaches a class on corpus linguistics, and the other has written numerous journal articles about corpus linguistics. For the validation, a questionnaire containing two sections was developed. The first section measured the quality of the created corpus along the two dimensions suggested by Lewis (1997). The second part of the questionnaire investigated the usefulness of the corpus for this specific academic purpose, using the three areas described by Keck and Kim (2014). Additionally, the questionnaire also inquired whether the inherent technical requirements of computer-based corpus usage outweigh the potential usefulness. After each question, additional room for comments or suggestions was provided.

All the experts agreed or strongly agreed that the corpus contained relevant jargon. All of the experts also agreed or strongly agreed that the corpus produced meaningful results when conducting legal English research. As discussed above, these two dimensions establish the overall validity of the corpus created for this project by confirming that it contains sufficient and diverse legal English jargon.

All experts agreed or strongly agreed that the corpus would provide information about word relationships that could be integrated into lesson planning. One commented “teachers can search for examples and actual uses of specific words or jargon, and use those examples to create texts or exercises for the class without actually interacting with the corpus in class.” However, the experts had concerns that familiarity with the software much be achieved first. For example, one expert stated “I strongly believe that users need to have a training session where they can get their hands on the tool before they can confidently use this tool.” Finally, all experts agreed regarding the

usefulness of the corpus as a self-study tool for learners, but thought it was better suited towards high-proficiency learners than low-proficiency. One expert lamented the lack of localization or an instruction manual in Thai for the AntConc program. A subsequent search of available online resources was unable to locate a version of AntConc with Thai menus or user documentation in Thai. Accordingly, a minimal level of English proficiency is a prerequisite for using AntConc for corpus analysis.

In conclusion, all three experts agreed that the corpus made for this project is useful to facilitate legal English learning. Therefore, the corpus has been found to possess relevant dimensions of quality, and found to be helpful for learning legal English.

**A guide to using the corpus**

Because AntConc may be non-intuitive to use, this section briefly describes analysis techniques available using the corpus. This section shows how to perform analyses via the Concordance, File View, Clusters/N-Gram, and Collocate tabs using the Florida – New York sub-corpus. Specific examples are given so that potential users may use this information as a guide on how to use the corpus, and how to select the methods best suited for various queries.

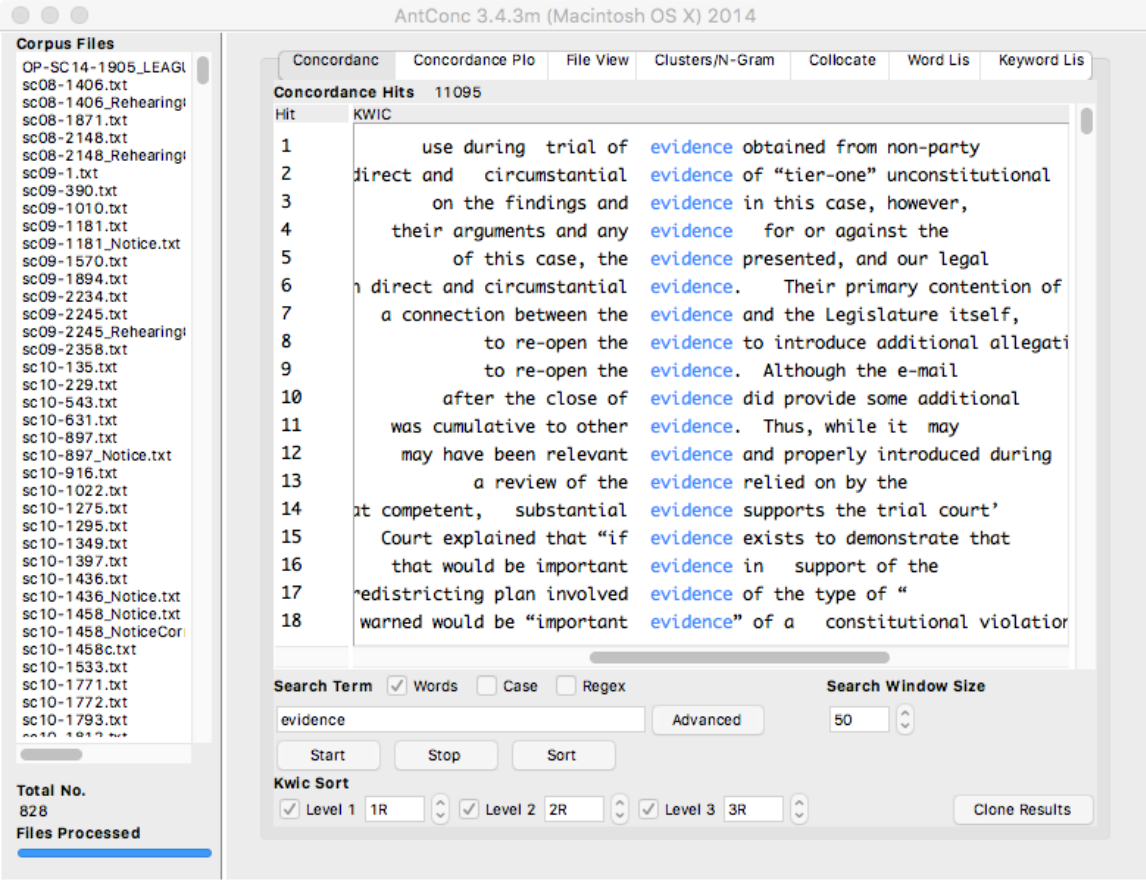
The first step is to select and open a corpus, which is accomplished by opening AntConc then the Open Dir… option in the File menu and selecting the folder containing the desired corpus. Additional corpora may be added to the AntConc analysis by repeating the same sequence. The major functions of AntConc may then be accessed by choosing the associated tab at the top of the window as shown in Figure 1.

***Concordance***

Most corpus analysis software, including AntConc, has a concordance function, wherein the corpus may be searched for keywords, which are then shown in context. After selecting the Concordance tab, entering a search query in the box towards the

bottom of the screen shows lexicon containing the query found in the corpus. Figure 1 shows the concordance search results obtained via searching for 'evidence.' Viewing just the first few lines of the concordance reveals modifiers for 'evidence,' such as circumstantial and relevant as well as action words, such as re-open. On larger computer monitors, this view can be expanded to show more text.

**Figure 1**

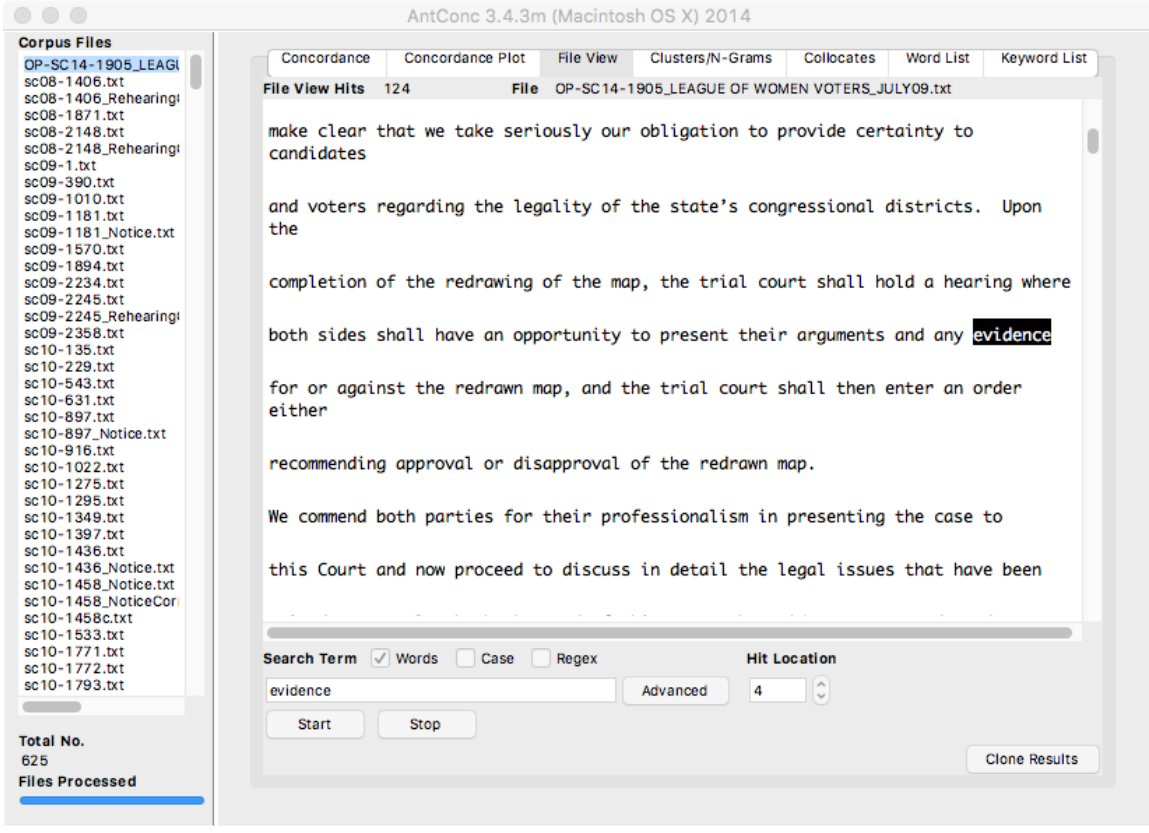



***File View***

Selecting any individual concordance will switch to the File View tab, showing the complete document and highlighting the keyword. Figure 2 is obtained by selecting the fourth concordance result from Figure 1, showing the full context of the keyword from the original document. Note how the selected keyword is highlighted in File View. The Hit Location box shows that the highlighted term is the fourth

instance of the term within this single document. Pressing the up or down arrows next to the number will skip forwards or backwards to other instances of the keyword.

**Figure 2**

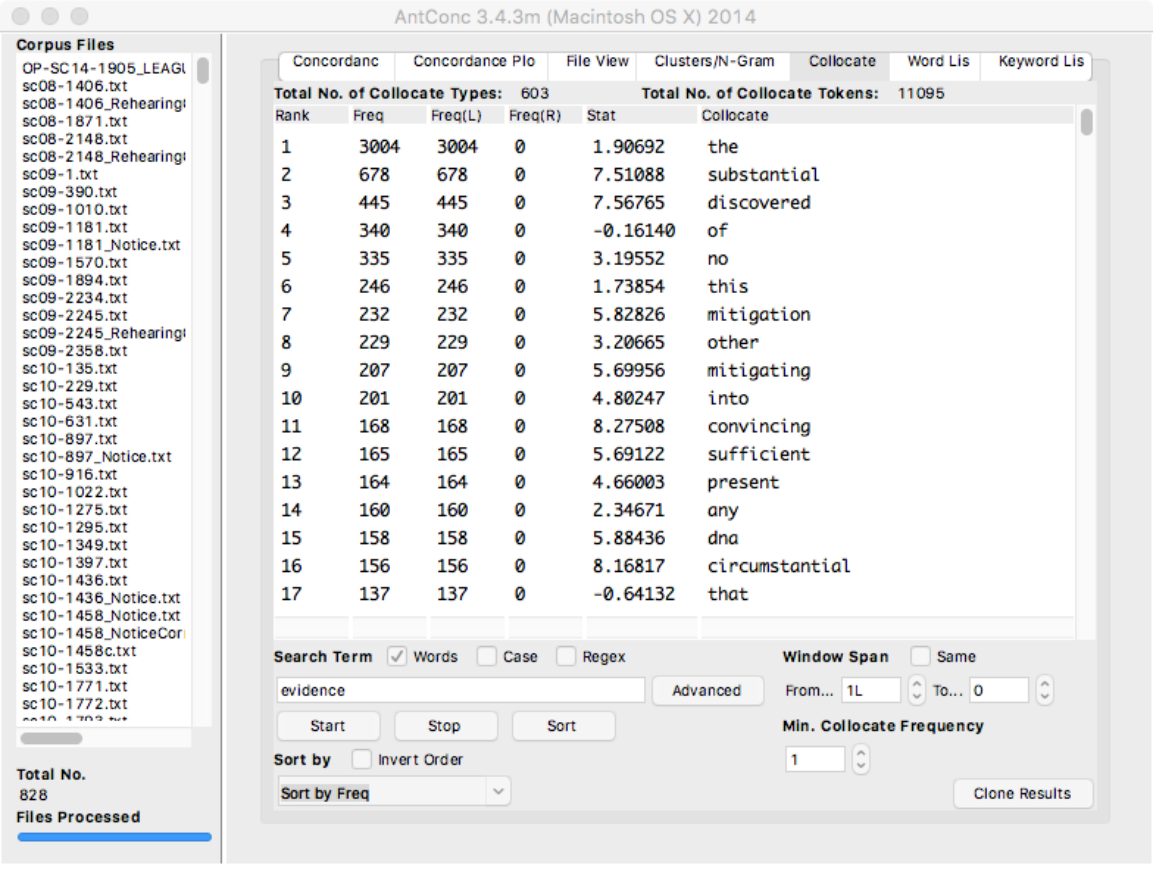



***Collocation***

Selecting the Collocates tab allows learners to search for words that frequently appear in close proximity to the search query. Figure 3 shows a search for collocates immediately preceding 'evidence'. In English, modifiers frequently appear in this position relative to the nouns they modify. Accordingly, the Figure 3 search results show modifiers commonly associated with the noun 'evidence' such as substantial, convincing, and circumstantial. For legal English analysis, ensure that Sort By Frequency at the bottom of the window. Note that Figure 3 represents merely the first screen of the search results, and many more results can be found by scrolling down the list.

Additionally, clicking on any collocate opens the selected phrase in the Concordance tab.

**Figure 3**

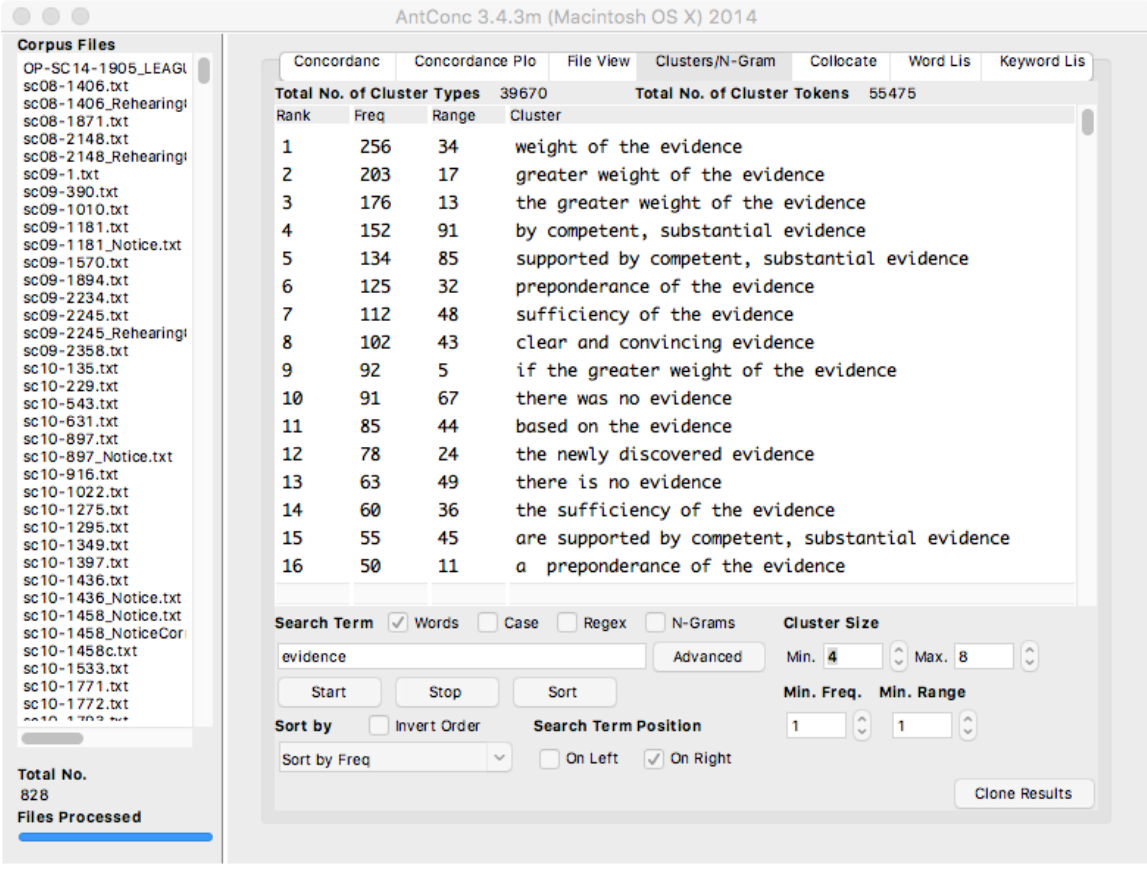



***Clusters / N-Gram***

Clusters are the commonly repeating chunks of language described by Lewis (1993). AntConc allows users to choose the size of the cluster, and also whether the search term appears as the first or last word in the cluster. Figure 4 shows the results of searching for clusters of four to eight words with 'evidence' in the final position. The first three results contain the idiom 'weight of the evidence.' Important legal standards of evidence, such as 'preponderance of the evidence' and 'clear and convincing evidence' also appear frequently in the corpus. Clicking on any cluster opens the Concordance tab for this specific chunk of language. The other function available via this tab, N-

Grams, does not use a search tem and provides a list of all chunks of a selected word length, sorted by frequency. Depending on the size of the corpus, the N-Grams function may take several minutes to complete.

**Figure 4**




***Other Features***

AntConc has three other functions, which will be briefly discussed here. First, the Word List tab provides a simple list of all words appearing in the corpus, sorted by frequency. Second, the Concordance Plot tab shows every location of a search query within each text using a graphic similar to a barcode. Finally, the Keyword List tab is not used by this project, as it requires additional resources to function.

**Discussion**

For this project, a sixteen million-word corpus was created using high-quality authentic legal texts specifically chosen to facilitate the study of legal English by non-native speakers. This project expands on prior works creating corpora tailored for specific academic purposes (O'Keeffe, McCarthy, & Carter, 2007) by constructing a corpus of this size to facilitate the study of legal English. Prior legal English corpus projects focused on translation (e.g Fan & Xunfeng, 2002), or used only concordance (e.g. Hafner & Candlin, 2007), rather than the advanced search functions of collocation and clustering available with AntConc. Therefore, this project unites corpus analysis with legal English learning. While other studies have discussed user-created corpora (e.g. Charles, 2012; Yoon & Hirvela, 2004), this project demonstrates how readily the process of corpus construction may be applied towards published American appellate court decisions to create a high quality corpus of legal English. Teachers may use the created corpus to find word relationships and set phrases to assist in lesson planning, and learners may use it to enhance communicative skills by locating set phrases, common modifiers, and action words.

The input from the expert evaluation provided useful insights into the strengths and weaknesses of the corpus. While the corpus is a powerful learning and teaching tool, optimal use requires requisite computer skills. This result was anticipated in accordance with the findings from prior studies (Hafner & Candlin, 2007). However, the lack of Thai localization or instructions for AntConc software may compound this issue for potential users of this project. Interestingly, most concerns of the experts were related to the technical nature of corpus analysis, rather than issues with the corpus created for this project.

Similar to prior studies about corpora designed for specific academic purposes (e.g. Charles, 2014; Hafner & Candalin, 2007), more advanced corpus construction techniques, such as tagging, were not used for this project. Because this corpus is freely available for educational purposes, any interested user may utilize parsing tools to

tag the part of speech for each word. AntConc can understand these tags, permitting learners to perform more advanced queries.

Meaningful further research may also be conducted that will build upon this project. First, making the software more accessible to Thai students by translating the software or manuals would be of great assistance to Thai users. Second, other users can use these same techniques to create their own legal English corpora from other jurisdictions, such as England or Hong Kong. This will enhance overall legal English learning by broadening the field of available corpora by encompassing local jargon from other jurisdictions. Due to the exclusive use of American legal materials for this corpus, searches for terms exclusive to other jurisdictions, such as 'barrister,' will not produce results. Hopefully, others who create legal English corpora will decide to make their corpora available free for educational use.

**Conclusion**

Using corpora to facilitate legal English learning is currently uncommon. However, this project shows that corpora may be useful tools for both teachers and learners of legal English. Making this corpus free for educational use will hopefully enhance worldwide learning of legal English. Interested users may download the complete corpus package by visiting https://goo.gl/51gcfn.

**The Authors**

Jason Skier is a master's degree candidate from the Teaching English as a Foreign Language Program (International Program) at the Faculty of Education, Chulalongkorn University.

Jutarat Vibulphol, PhD., is an instructor at Division of Foreign Language Teaching, Faculty of Education, Chulalongkorn University, Bangkok, Thailand.

**References**

Anthony, L. (2014). *AntConc Homepage*. Retrieved 10 15, 2015, from Lawrence Anthony's Website: http://www.laurenceanthony.net/software/antconc/

Anthony, L. (2016). *AntFileConverter homepage*. Retrieved 5 17, 2016, from Lawrence Anthony's Website: http://www.laurenceanthony.net/software/antfileconverter/

Charles, M. (2014). Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes* , 30-40.

Collins. (2013). *The History of COBUILD*. Retrieved 10 15, 2015, from www.collins.co.uk: http://www.collins.co.uk/page/The+History+of+COBUILD

Coxhead, A. (1998). *An academic word list (Vol. 18).* School of Linguistics and Applied Language Studies, Victoria University of Wellington.

Crosthwaite, P. (2012). Learner corpus linguistics in the EFL classroom. *PASAA Journal*, *44*, 133-147.

Csomay, E., & Petrović, M. (2012). "Yes, your honor!": A corpus-based study of technical vocabulary in discipline-related movies and TV shows. *System , 40* (2), 305-315.

Fan, M., & Xunfeng, X. (2002). An evaluation of an online bilingual corpus for the self-learning of legal English. *System , 30* (1), 47-63.

Hafner, C., & Candlin, C. (2007). Corpus tools as an affordance to learning in professional legal education. *Journal of English for academic purposes , 6* (4), 303-318.

*Hawaii Rules of Appellate Procedure 28(b)(7).* (1984). Retrieved 10 15, 2015, from Hawaii Rules of Appellate Procedure: http://www.courts.state.hi.us/docs/court_rules/rules/hrap.htm

Keck, C., & Kim, Y. (2014). *Pedagogical grammar.* John Benjamins Publishing Company.

Lewis, M. (1997). *Implementing the lexical approach: Putting theory into practice.* London: Commercial Colour Press.

Lewis, M. (2001). *Teaching collocation: Further developments in the lexical approach.* London: Commercial Colour Press.

Lewis, M. (1993). *The lexical approach.* Hove: Language Teaching Publications.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching.* Cambridge University Press.

State v. Haanio, 94 Haw. 405, 415, 16 P.3d 246, 256 (2001).

Stith, K. (1990). The Risk of Legal Error in Criminal Cases: Some Consequences of the Asymmetry in the Right to Appeal. *The University of Chicago Law Review , 57* (1), 1-61.

Yoon, H., & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of second language writing , 13* (4), 257-283.

## Appendix A

Corpus Contents

| Court | Year(s) | # Documents | Word Tokens | Word Types |
|---|---|---|---|---|
| U.S. Supreme Court | 1997, 1998, 1999, 2013, 2014 | 1,168 | 5,370,134 | 50,563 |
| U.S. Supreme Court Transcripts | 2012, 2013, 2014 | 138 | 1,776,879 | 19,587 |
| Hawaii | 2014, 2015 | 2,563 | 4,866,747 | 32,314 |
| Florida-New York | 2013, 2014, 2015 | 828 | 4,214,379 | 33,350 |
| TOTAL | | 4,697 | 16,229,139 | 75,237 |