EuroAmerican Journal of Applied Linguistics and Languages
E-JournALL Volume 1, Issue 1, November 2014, pages 48-68
ISSN 2376-905X
DOI http://dx.doi.org/10.21283/2376905X.1.4
www.e-journall.org

Exploring mega-corpora: Google Ngram Viewer and the Corpus of Historical American English

ERIC FRIGINAL[a1], **MARSHA WALKER**[b], **JANET BETH RANDALL**[c]
[a]Department of Applied Linguistics and ESL, Georgia State University
[b]Language Institute, Georgia Institute of Technology
[c]American Language Institute, New York University, Tokyo

Received 10 December 2013; received in revised form 17 May 2014; accepted 8 August 2014

ABSTRACT

EN The creation of internet-based mega-corpora such as the Corpus of Contemporary American English (COCA), the Corpus of Historical American English (COHA) (Davies, 2011a) and the Google Ngram Viewer (Cohen, 2010) signals a new phase in corpus-based research that provides both novice and expert researchers immediate access to a variety of online texts and time-coded data. This paper explores the applications of these corpora in the analysis of academic word lists, in particular, Coxhead's (2000) Academic Word List (AWL). Coxhead (2011) has called for further research on the AWL with larger corpora, noting that learners' use of academic vocabulary needs to address for the AWL to be useful in various contexts. Results show that words on the AWL are declining in overall frequency from 1990 to the present. Implications about the AWL and future directions in corpus-based research utilizing mega-corpora are discussed.

Keywords: GOOGLE N-GRAM VIEWER, CORPUS OF HISTORICAL AMERICAN ENGLISH, MEGA-CORPORA, TREND STUDIES.

ES La creación de megacorpus basados en Internet, tales como el Corpus of Contemporary American English (COCA), el Corpus of Historical American English (COHA) (Davies, 2011a) y el Visor de Ngramas de Google (Cohen, 2010), anuncian una nueva fase en la investigación basada en corpus, pues proporcionan, tanto a investigadores noveles como a expertos, un acceso inmediato a una gran diversidad de textos online y datos codificados con *time-code*. Este artículo explora las aplicaciones de estos corpus en el análisis de listas de vocabulario académico, en particular, Coxhead's (2000) Academic Word List (AWL). Coxhead (2011) hizo patente la necesidad de seguir investigando las aplicaciones del AWL con corpus más amplios, al apuntar a que el uso de vocabulario académico por parte de los aprendices necesita ser considerado para que el AWL sea útil en diferentes contextos. Los resultados muestran que la frecuencia de uso general de las palabras contenidas en el AWL está disminuyendo desde 1990. Asimismo, se tratan los efectos de esta tendencia en el AWL y las futuras líneas de investigación de estudios que utilizan megacorpus.

Palabras clave: VISOR DE NGRAMAS DE GOOGLE, CORPUS HISTÓRICO DE INGLÉS AMERICANO, MEGACORPUS, ESTUDIOS DE TENDENCIAS

IT La creazione di mega-corpora basati su internet, come il Corpus of Contemporary American English (COCA), il Corpus of Historical American English (COHA) (Davies, 2011a) e il Google Ngram Viewer (Cohen, 2010), inaugura una nuova fase della ricerca basata su corpora che permette a ricercatori e ricercatrici junior e senior di accedere a un'ampia gamma di testi on-line e dati con codifica *time-code*. L'articolo esplora le applicazioni di questi corpora nell'analisi di glossari (*word lists*) accademici, in particolare, l'Academic Word List (AWL) di Coxhead (2000). Coxhead (2011) ha sollecitato ulteriori ricerche sull'AWL tramite corpora di estensione maggiore, sottolineando la necessità di considerare l'uso del lessico accademico da parte degli apprendenti affinché l'AWL sia utile in molteplici contesti. I risultati mostrano che dal 1990 a oggi la frequenza generale dei lemmi dell'AWL è diminuita. Seguono considerazioni sulle implicazioni riguardanti l'AWL e i futuri orientamenti della ricerca basata sull'uso di mega-corpora.

Parole-chiave: GOOGLE N-GRAM VIEWER, CORPUS OF HISTORICAL AMERICAN ENGLISH, MEGA-CORPORA, STUDI DI TENDENZA.

[1] Contact: efriginal@gsu.edu

1. Introduction

Over the past several years, research in corpus linguistics has produced generalizable and accurate linguistic information that has been useful in analyzing variation in language (Biber, Conrad, & Reppen, 1998). Enabled by major advancements in computational technology, software design, and the internet, corpus-based approaches have continued to explore specialized corpora and extensive applications of quantitative data which would otherwise be infeasible in traditional research methodologies. As a result, researchers using corpora are able to pursue a dynamic set of research questions that often result in radically different perspectives on language variation and use from those taken in previous research (Biber, Reppen, & Friginal, 2010). In addition, the number of corpora freely distributed online has increased tremendously, allowing both novice and expert researchers the opportunity to obtain relevant linguistic frequency and distributional data immediately. Searchable corpora and databases with built-in interfaces that provide part-of-speech (POS) tags, demographic information of writers and speakers of texts, and register-specific comparative charts can now easily be retrieved online. This opportunity has made the domain of web-based, corpus research user-friendly and accessible. Developments such as these have redefined the nature of linguistic research with corpora, which has traditionally been associated only with a specialized group of linguists in academic settings.

Biber, Conrad, and Reppen (1998) define "corpus" and "corpus representativeness" as:

> […] not only a collection of texts. Rather, a corpus seeks to represent a language or some part of a language. The appropriate design for a corpus therefore depends upon what it is meant to represent. The representativeness of a corpus, in turn, determines the kind of research questions that can be addressed and the generalizability of the results of the research. (p. 246)

This definition and operationalization of a corpus and its inherent representativeness were especially relevant during pre-computer corpus linguistics and were broadly emphasized in earlier periods of computer-based corpus design, such as during the development of the Brown corpus and its cognates, known as the first family of corpora (Svartvik, 2007). Researchers have debated the need to delineate how large a corpus must be in order to accurately represent the language of a particular population of writers/speakers or the distribution of certain grammatical features (e.g., Biber, 1990, 1993; McEnery, Xiao, & Tono, 2006; Miller, 2011). The corpus-driven approach, which prescribes formulating an overarching research question as a primary step before initiating the collection of a corpus, follows corpus design principles and standards based on the concepts of representativeness and generalizability of data and anticipated results (McEnery, Xiao, & Tono, 2006). At the same time, however, texts intended to represent varieties and sub-varieties of languages have been compiled over the years, producing large-scale general corpora, such as the British National Corpus (BNC) and the American National Corpus (ANC), which can be said to represent two national varieties of English, and the International Corpus of English (ICE), which purportedly represents international/global varieties of educated English (Bautista, 2011; Nelson, 1996). These three corpora will be discussed further in the sections below.

The successful collection of internet-based mega-corpora (>300 million words) such as the Corpus of Contemporary American English (COCA), the Corpus of Historical American English (COHA) (Davies 2005, 2011a, 2011b), and the Google Ngram Viewer (Cohen, 2010; Hotz, 2010; Michel, Shen, Aiden, Veres, & Gray, 2011) may signal a new phase in corpus design in which sampling texts in order to achieve statistical representativeness will no longer be required. Mega-corpora have the potential to cover an entire population of texts in a particular register instead of only presenting a sample, as in the current efforts of Google Books to scan and compile books and manuscripts produced since the 1500s across many languages (Toor, 2010). The clear message here is that technology will eventually enable researchers to have easy access to most, if not all, texts in a register. For example, access to all internet-based texts (e.g., social media language through Facebook and Twitter posts, personal blogs, news reports, etc.) could be entirely and automatically obtained without the need for sampling or manual/selective crawling.

1.1. Analysis of large electronic corpora

Studies based on large electronic corpora began in the late-1960s with the Brown Corpus, a one million word collection of published written texts in American English, compiled by Kucera and Francis. A parallel corpus of British written texts, the Lancaster-Oslo-Bergen (LOB) Corpus was published subsequently.

Kucera and Francis pioneered extensive work on word frequencies and distributions of grammatical part-of-speech categories using the Brown and LOB corpora (Biber, 1995). In the early 1980's, the increasing availability of personal computers, electronic corpora, and corpus tools such as concordancers, taggers, and parsers facilitated wide-ranging linguistic analyses. In those years, descriptive studies featuring the distributions of linguistic features (e.g., passives, verbs, formulaic sequences, nominalizations) across corpora were conducted and published. During the same period, dictionaries, such as the Collins COBUILD English Language Dictionary (1987) and the Longman Dictionary of Contemporary English (1987), intended for language learners and based on the analysis of large electronic corpora, began to be published (Biber, Reppen, & Friginal, 2010).

From the late 1980's, Biber developed quantitative studies of tagged texts using advanced multivariate statistical techniques (e.g., Factor and Cluster Analyses) and focusing on the concept of statistical co-occurrence of linguistic features across registers. Biber's comparison of large volumes of texts representing sub-registers of speech and writing in cross-linguistic settings provided new approaches in simultaneous explorations of general and specialized corpora (Friginal, 2009). The Longman Grammar of Spoken and Written English (Biber, Johansson, Leech, Conrad, & Finegan, 1999), for example, provided wide-ranging distributional comparisons of linguistic features from spoken and written British and American Englishes. These descriptive data on lexico-grammatical variation have also been used in the design of classroom activities for English language teaching, especially for non-native speakers of English.

Exponential developments in internet technology in the last decade have positively altered the nature of corpus-based research. Collecting written texts such as newspaper articles, published academic papers, opinion columns, and weblogs became easily manageable, as more and more types of specialized texts have been uploaded and distributed online (Grieve, Biber, Friginal, & Nekrasova, 2010; Herring & Paolillo, 2006). Although there are still many challenges in the collection of various specialized corpora, especially in the domain of spoken discourse, the current set of freely available online corpora shows how corpus linguistics has progressed from the 1960s to the present (Preiss, Coonce, & Baker, 2009). The internet as a corpus (Crystal, 2006) not only covers English texts but also other languages used online by an increasing number of users. Automated transcriptions of online audio and video clips and translation services (e.g., Dragon Dictation software, Google Translate from Google Labs) have also been introduced and developed, showing promising applications that could help in the collection of spoken texts. These online tools greatly contribute to how corpora are now collected and may necessitate new models in corpus design and compilation.

1.2. Development of English electronic mega-corpora

Since its release in the early 1990s, the 100 million word British National Corpus (BNC, including recent XML version) has been used in more corpus-related studies than any other English-language corpus (Davies, 2009). The BNC, with its web-based interface, BNCweb, (http://bncweb.info/) provides a client-program for searching and retrieving lexico-syntactic and textual data from available metadata of writers'/speakers' demographic information and registers. The BNC served as the model for the creation of the American English parallel corpus, ANC (http://americannationalcorpus.org/), which has released two versions of component sub-corpora totaling over 22 million words. Work on the ANC is has progressed slowly and there is currently no user-friendly client program available for automatic database online searches. However, the ANC has freely downloadable texts totaling over 15 million words and the annotated versions of the corpus also include grammatically-tagged data and other XML annotations.

Davies (2009, p. 59) pointed out that, valuable as it is, the BNC "is beginning to show its age in some respects." He noted that while files and demographic information on some registers have been corrected or updated, substantive additions to the corpus have been very limited since 1993. Because there is no currently planned expansion to the BNC, it may become increasingly out of date with respect to recent changes in English and the need to represent additional registers such as texting/SMS language and the online language of social media and blogging. Similarly, the ANC's collection of texts reflects work done from early- to mid-2000s with no continuing contributions that allow for diachronic comparisons across registers.

The Corpus of Contemporary American English (COCA), developed and published online by Davies, was released in early 2008. It covers a diverse collection of American English texts totaling more than 385 million words from 1990–2008 (20 million words each year) across registers grouped into the following categories: spoken, fiction, popular magazines, newspapers, and academic journals. Since its release, COCA's online interface has been widely used by researchers, teachers, and students for various purposes, including

producing materials for teaching collocations, lexico-syntactic features of English, and diachronic word frequency changes across registers. COCA is comparable to the BNC/ANC in terms of text types, though it deviates from them in the types of spoken data it includes: COCA's spoken texts (20% of the corpus) come mainly from television news and interview programs, rather than from the types of conversation data (e.g., face to face conversation, service encounters, and telephone interactions) available in the BNC or other corpora such as the Longman Corpus. Davies (2009) maintains, however, that COCA's overall balanced composition means that researchers can compare data across registers and achieve relatively accurate results that show patterns of change in the language from the 1990s to the present.

In addition to these three corpora, the International Corpus of English (ICE) has also been created and made available since the mid-1990s. The ICE consists of one million words per spoken and written variety of English produced by each of over 20 research teams worldwide (http://ice-corpora.net/ice/). For most participating countries, the ICE project served as the first systematic investigation of the national, "educated" English variety (Nelson, 1996). Each component corpus follows a common corpus design and a similar scheme for grammatical annotation, as the ICE was primarily intended for comparative studies of emerging Englishes all over the world. For example, Asian varieties of English available for free download from the ICE website include sub-corpora from countries/territories such as Hong Kong, the Philippines, India, and Singapore, where English has been used extensively as the language of business and education. Written registers in the ICE range from student writing, novels and stories, social and business letters, to published, professional texts. The spoken data component features face-to-face conversations, broadcast speech, spontaneous commentaries, telephone calls, parliamentary debates, and legal cross-examinations, among other groups of texts.

Finally, and as the primary focus of this paper, the internet has paved the way for Google Ngram Viewer and COHA, two large online databases of, specialized categories of texts, released in 2010, which reflect current academic and corporate efforts to freely distribute corpus-based data to the public. These two mega-corpora are both designed to address diachronic analyses, but can also be used for synchronic studies due to the size of data representing recent published texts (Davies, 2011a). Both of these applications aim at continuing to periodically add texts to their databases. A number of studies exploring these corpora's potential usefulness in language analysis and corpus linguistics were initiated in the months after they became publicly available; a review of three recent conferences, including the 2011 and 2013 American Association for Corpus Linguistics Conference, shows that over 25 papers were presented using either COHA or Google Ngram Viewer.

2. Exploring mega-corpora

2.1. Google Ngram Viewer

Google Ngram Viewer is comprised of over 5.2 million books from Google Books, a subdivision of Google Inc. that has conducted an extensive scanning of published manuscripts in order to create a database of electronic or digitized texts. The number of currently scanned books comprises approximately 4% of all the books ever written in English (Bohannon, 2011). This mega-corpus contains over 500 billion words; the majority of them are in English (361 billion). Other available languages in the corpus include French, Spanish, German, Chinese, Russian, and Hebrew (Bohannon, 2011; Keuleers, Brysbaert, & New, 2011; Michel et al., 2011). The Ngram Viewer contains data that span from 1550-2008, although texts before the 1800s are extremely limited and often there are only a few books per year. From the 1800s, the corpus grows to 98 million words per year; by the 1900s, it reaches 1.8 billion, and 11 billion per year by the 2000s (Michel et al., 2011). Hence, texts collected for Google Ngram Viewer represent the largest corpus to date.

Google Ngram Viewer is composed of raw data that is encoded by the number of n-grams, adjacent sequences of n items from a text. The n-gram was developed as a response to the concerns of Peter Norvig, the head of research at Google Labs, about developing the online viewer interface due to many pending lawsuits about Google's book digitizing initiative (Bohannon, 2010). In order to avoid further problems with publishers and copyright owners of published materials, the Ngram Viewer makes use of texts converted into various n-grams so that data from the whole corpus cannot be downloaded as complete books, and in effect, "cannot be read by a human" (Michel et al., 2011, p.176). The process of converting Google Books into n-grams has been tedious, and, as result, just 5.2 million of the 15 million Google books have been converted so far. Google's conversion procedure is as follows:

- Raw texts extracted from different sources are pre-processed by tokenizing white-spaces; upper case letters are converted to lower case.
- Numbers are retained, and no stemming/inflection are performed. The n-gram LMs (Language Models) are word-based backoff models, where the n-gram probabilities are estimated using Maximum Likelihood Estimation (MLE) with smoothing (Gao, Nguyen, Li, Thrasher, Li, & Wang, 2010).
- The viewer is composed of n-grams, from 1-gram to 5-grams. N-grams represent how many words are in a lexical bundle: 1-gram = 1 word, while 5-grams = 5 words (Bohannon, 2011). Once the n-grams are saved as raw data in Excel files, the n-grams can then be searched using the online viewer.
- Search results are displayed by frequency for each n-gram determined per year following the formula: total number of instances of the n-gram / total number of words in that year.

It is important to note that Google Ngram Viewer was not created with research in linguistics or corpus linguistics as its primary application (Michel et al., 2011). The developers of the viewer wanted to create a new approach to humanities research that they coined Culturomics. Culturomics (www.culturomics.org/home) is a way to quantify culture by analyzing the growth, change, and decline of published words over centuries. This would make it possible to rigorously study the evolution of culture using distributional, quantitative data on a grand scale (Bohannon, 2010). In an effort to prove the adequacy of and to provide clear impetus for Google Ngram Viewer, Google-affiliated researchers have conducted a series of studies to validate the usefulness and various applications of their new program. One study, for example, showed that over 50% of the words in the n-gram database do not appear in any published dictionary (Bohannon , 2010). In addition, it is argued that patterns and cultural influences of words could be clearly established and tracked across timeframes. The use of Google Ngram Viewer and Culturomics, therefore, contributes academic and technical value to the study of culture, making it arguably a new cultural tool that has several possibilities. Figure 1 displays the default search screen of Google Ngram Viewer as of September 2014.

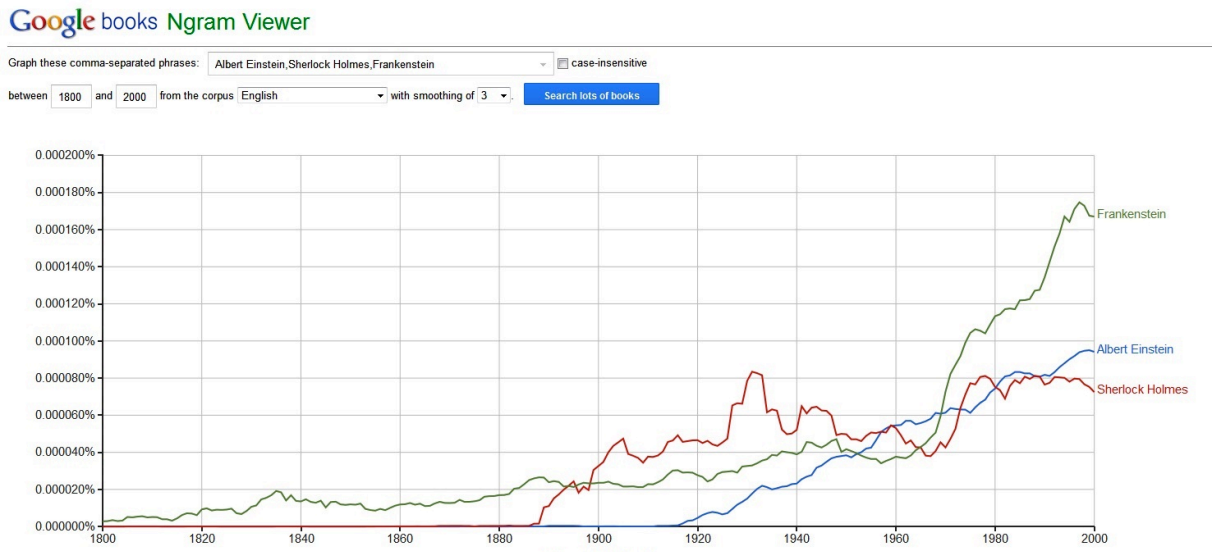

Figure 1. Default search screen of Google Ngram Viewer (http://books.google.com/ngrams)

2.2. Corpus of Historical American English (COHA)

Like the Google Ngram Viewer, the Corpus of Historical American English was debuted in 2010 with limited functionality. Information available from COHA has been continually updated since then with major upgrades in its data and search features, including interesting information on visualization techniques as applied to COHA's diachronic distributional data (Hilpert, 2011). COHA was developed by Mark Davies, Professor of Corpus Linguistics, from Brigham Young University. As noted earlier, Davies also created the

Corpus of Contemporary American English and many other interactive online databases for corpus analyses, such as the Corpus do Português, Corpus del Español, and Corpus of LDS General Conference Talks, which can be accessed from his personal website (http://davies-linguistics.byu.edu/personal/). COHA was funded by the National Endowment for the Humanities as part of its "We the People" initiative (Davies, 2011a).

COHA has just over 400 million words, compared to Google Ngram Viewer's 500 *billion* words. Davies argues, however, that this substantial difference in corpus size does not necessary affect reliability of results when these two corpora are used and compared across a range of linguistic distributions. When searches were performed between COHA and Google Ngram Viewer, similar results were found once data were *normalized* (Davies, 2011a, 2011b). To date, COHA's data cover the period from 1810 to 2010, and specific registers such as fiction, magazine, non-fiction, and newspaper texts. Following the functionalities established by COCA, COHA also provides POS lists, collocates, list and chart options, and various sorting and comparison options. Figure 2 displays the default search screen of COHA.

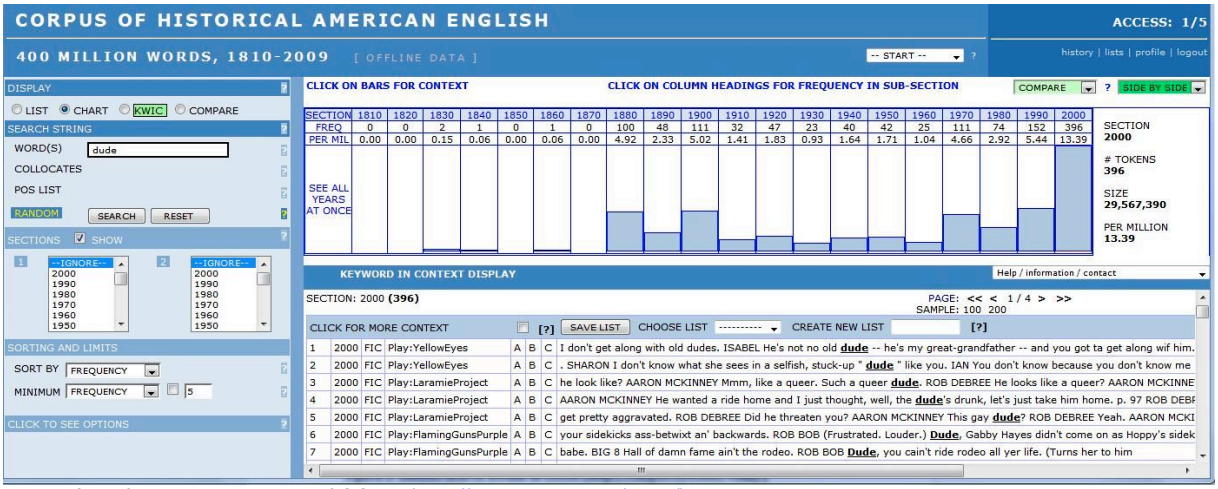

Figure 2. Default search screen of COHA (http://corpus.byu.edu/coha/)

According to Davies, COHA can do exactly what Google Ngram Viewer does, with the addition of specific features that directly apply to linguistics and corpus-based research. This includes the ability to employ concordancing tools, display data in context, and limit results by register, characteristics not available on Google Ngram Viewer. In addition, COHA allows more powerful searches, for example, for synonyms and word comparisons across historical periods - which may produce insightful analyses about cultural and societal shifts (Davies, 2011a). The types of detailed COHA searches include:

- lexis, through mass comparison between historical periods,
- morphology, via wildcards,
- syntax, from POS-tagged data included in the program, and
- semantics, via collocates, synonyms, and customized lists.

2.3. Davies' Google Books Interface (2011c)

In mid-2011, Davies launched an early version of an interface for the Google Ngram Viewer database that allows a more extensive search of the massive Google collection than the basic Ngram Viewer. While Davies points out that his interface is not an official product of Google or Google Books (Davies, 2011c), this search platform is very similar to the standard COCA/COHA structure providing options for wildcard, lemma, part of speech, synonyms, and collocates searches. This interface is more advanced than Google's online viewer, with presentations of frequency data that go beyond simple line graphs or figures. In addition, Davies' interface allows users to copy the data to other applications for further analysis (Davies, 2011c), a feature that is not currently possible in the Google interface. The Davies version is based on 155 billion words of American English from 1810-2009, with future plans to integrate other Google Books collections into this new interface, including texts from British and American English texts from the 1500s-1700s, as well as texts from Spanish, German, and French (a grant application to support this endeavor was submitted in mid-2011). Davies' interface also provides a guided tour of the site (http://googlebooks.byu.edu/) showing its major

features and an intuitive search form. A click for each search query automatically fills in the form and displays the results obtained from actual American English books collected by Google Books.

2.4. Recent linguistic analysis of data from COHA and Google Ngram Viewer

One way to explore mega-corpora such as Google Ngram Viewer and COHA is to conduct diachronic and trend studies comparing the distribution of various linguistic features across specific time periods. Davies (2011a), for example, presents frequency changes in lexical items (e.g., fellow, teenager, sublime, global warming, steamship) used in published texts from 1810 to the present. Hilpert (2011) uses visualization techniques using Google Visualization (or Google Code) and the statistical/graphical software, R, to show language change through COHA. Sociolinguistic time series studies on semantic shifts (semantic changes) across its designated registers can also be conducted using COHA, although this could be improved if corpora were to include demographic information about speakers/writers. While more generic trending information is provided by Google Ngram Viewer from published books and manuscripts, such a dataset also allows for immediate macro-level snapshots of quantitative language patterns across time, from the 1800s to the present.

Specialized linguistic work from Google Ngram Viewer has to date focused on cultural trends throughout history, seen through the books that have been scanned by Google. The Culturomics team looks quantitatively at linguistic and cultural phenomena reflected in the language of these texts. This reporting of linguistic trends is directly related to the fields of "lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology" (Michel et al., 2011, p.176). Some lexical and culturally-specific data intended for popular reading obtained from the Google Ngram Viewer and Culturomics include the following (from Cohen, 2010):

- women in comparison with men is rarely mentioned until the early 1970s, when feminism gained a foothold. The two topic lines eventually cross paths about 1986.
- Mickey Mouse and Marilyn Monroe do not get nearly as much attention in books as Jimmy Carter.
- There are many more references in English than in Chinese to Tiananmen Square after 1989.
- Grilling has been used frequently from the late 1990s and outpaced roasting and frying in 2004.

Michel et al. (2011) also measured the endurance of fame from the 1800s, reporting that written references to popular celebrities faded twice as quickly in the mid-20th century as they did in the early 19th. They also found technological advances and inventions (e.g., telephone, radio) took, on average, 66 years to be adopted by the larger culture in the early 1800s, and only 27 years between 1880 and 1920. They tracked the way irregular English verbs that did not add -ed at the end for past tense (i.e., learnt) evolved to conform to the common pattern (learned). And finally, as an application of quantitative, lexico-syntactic data, Michel et al. estimated that the English lexicon has grown by about 70 percent to more than a million words in the last 50 years with direct implications to lexicographic changes and updating dictionaries by pinpointing newly popular words and obsolete ones (Cohen, 2010).

3. More research with mega-corpora

3.1. Focus of the present case study

There are, however, possible lines of research using mega-corpora that move beyond culture-based analyses and diachronic studies. The remainder of this paper is dedicated to an exploratory case study that shows the applications of Google Ngram Viewer and COHA in the analysis of academic word lists, in particular, Coxhead's (2000) Academic Word List (AWL). The goal is not to compare these mega-corpora, but to add to the growing number of studies that explore the applications of new data sources that may serve as models for more specialized analyses of COHA and Google Ngram Viewer.

Word lists are a conventional concept common in corpus-based studies with emphasis on pedagogical applications of corpus data especially in the teaching of academic writing in English. The study of vocabulary use predates other areas of corpus investigation (Biber, Conrad, & Reppen, 1998) and vocabulary teaching materials developed from corpora have been commonly used in many writing classrooms and incorporated in dictionaries and textbooks (McEnery, Xiao, & Tono, 2006). The AWL has been widely used in English for Academic Purposes (EAP) and the teaching of second language writing in English (Reppen, 2010). Coxhead (2000, 2011) has called for further research on the AWL with larger corpora, noting that there is a

continuing need to address learners' use of academic vocabulary for the AWL to be useful in various contexts over a decade after its publication. Google Ngram Viewer and COHA could provide this opportunity for an update, as these mega-corpora have the volume of words and a variety of registers needed to check actual patterns of lexical distributions.

3.2. The Academic Word List

The AWL was created by Averil Coxhead, School of Linguistics and Applied Language Studies at Victoria University of Wellington, as a rationalized, more specialized extension of, or perhaps response to, the General Service List (GSL). The GSL is a list of over 2000 words that was developed by Michael West in 1953 and updated by John Bauman and Brent Culligan in 1995. The list includes the most frequent words of written English collected primarily for English language learners and ESL writing teachers. The updated version of the GSL uses frequencies from the Brown Corpus (Bauman, 1995). Words from the AWL came instead from an "Academic Corpus" that Coxhead collected herself. The corpus contains a total of 3.5 million words with texts representing multiple academic sources such as journals and textbooks published from 1993 to 1996. Coxhead also included texts from the Brown Corpus, London/Oslo/Bergen Corpus (LOB), and MicroConcord Academic Corpus. The four predominant academic registers include Arts, Commerce, Law, and Science, and encompass 28 different sub-registers (Coxhead, 2000, 2002).

An automated search of the Academic Corpus yielded the words comprising the AWL, which consists of 570 word families, and 3,110 individual words. The words were selected based on their specialized occurrence, range, and frequency, and only included if not already present in the first 2,000 words of the GSL. The total number of words on the list results in 10% of all the words in academic texts. As a rule, each word had to be used at least 100 times in the Academic Corpus. Table 1 shows sample entries from the AWL with each lemma listed on the left and all of its lexemes listed underneath each lemma. The bolded words were the most frequent on the sub-list. The most frequent word in each word family could be the lemma (like the word estimate seen below) or it could be one of the lexemes (like the word derived also below). The AWL has 10 sublists, with Sublist 1 containing the highest frequency words.

Table 1
Sample sublist from AWL

derive	**estimate**	**function**
derivation	estimated	functional
derivations	estimating	functionally
derivative	estimation	functioned
derivatives	estimations	functioning
derived	over-estimate	functions
derives	overestimate	
deriving	overestimated	identify
	overestimates	identifiable
distribute	overestimating	identification
distributed	underestimate	**identified**
distributing	underestimated	identifies
distribution	underestimates	identifying
distributional	underestimating	identities
		identity

As noted above, Coxhead's primary reason for creating the AWL was that other word lists, especially the GSL, were limited in their capacity to demonstrate current lexical usage across academic registers. Yet, after more than 10 years, it is possible that the AWL may also need some updating. For example, Ming-Tzu and Nation (2004) completed a study on homographs within the AWL and concluded that the list should include a wider range of word families and lemmas, representing a range of academic homographs. In addition, Nation and Waring (1997) also argued that in order to comprehend a text, 95% to 98% of words in the text must be fully understood and acquired by learners. Unfortunately, as Nation and Waring found, word lists such as the AWL and GSL arguably do not represent at least 95% of words in a target text.

3.3. Data collection

The overarching research question of this exploratory case study is: How are words on the AWL reflected and distributed across recent time frames in Google Ngram Viewer and COHA? In pursuing this line of research, it may also be possible to answer whether or not there is evidence that an updated word list is necessary. The words analyzed in the study are from AWL Sublist 1. This list contains 60 words and their word families. The most frequent word from each lemma was selected for frequency searches in the two mega-corpora (see Appendix for the complete list of 60 words used).

Google Ngram Viewer was searched to find words with "increase or decrease" counts from 1998 to 2008 that are greater than or equal to .0050. From these results, 10 words were chosen to be further analyzed qualitatively. One of the ten words, for example, was *labour*, which was found to have a rapid decrease in the given time frame. However, since *labour* is typically British English and not American English, it was excluded from this case study (as COHA represents only American English). Because of that exclusion, nine words were ultimately selected for distributional analysis using Google Ngram Viewer: *analysis, data, economic, environment, policy, research, section, structure*, and *theory*. COHA, from the same time frame, was used for *economic, evidence, individual, major, percent, period, principle, section*, and *theory*. Words with frequencies with a difference greater than or equal to 20 tokens per million were selected for additional analysis. This was done as COHA lists actual and normalized frequencies compared to percentages from Google Ngram Viewer.

4. Results

4.1. Macro-level distributions of 15 AWL words from Google Ngram Viewer

Of the 60 AWL words analyzed in this study for both platforms, more than 90% are shown to have declined in use from the late-1990s to the present. Fifteen words from AWL were found to have higher levels of differences from the set of results obtained from Google Ngram Viewer and COHA. These are: (1) analysis, (2) data, (3) economic, (4) environment, (5) evidence, (6) individual, (7) major, (8) percent, (9) period, (10) policy, (11) principle, (12) research, (13) section, (14) structure, and (15) theory. Figure 3 shows Google Ngram Viewer's word chart tracking and distributions of some AWL words from the 1900s to 2000s.

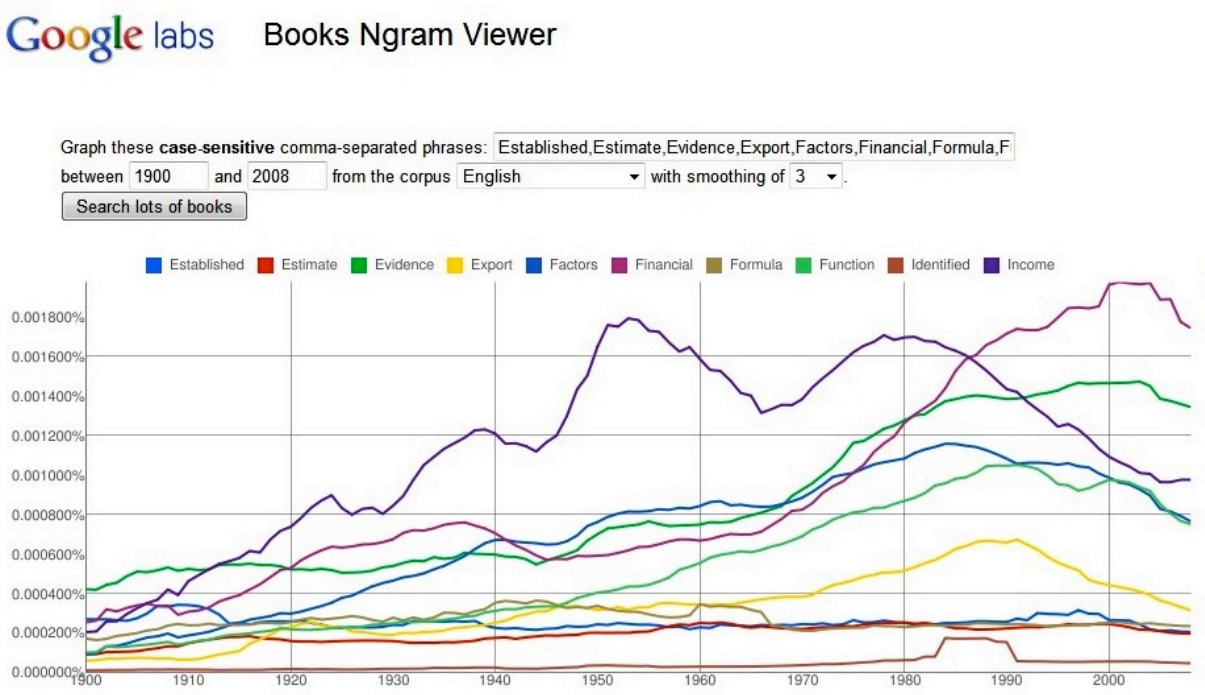

Figure 3. Sample distributional chart tracking of AWL words from Google Ngram Viewer, 1900s to 2000s

Specifically, for the 15 words listed above, Figures 4 and 5 show a slow decline in their general frequencies from English books from 1998 to 2008 from Google Ngram Viewer.

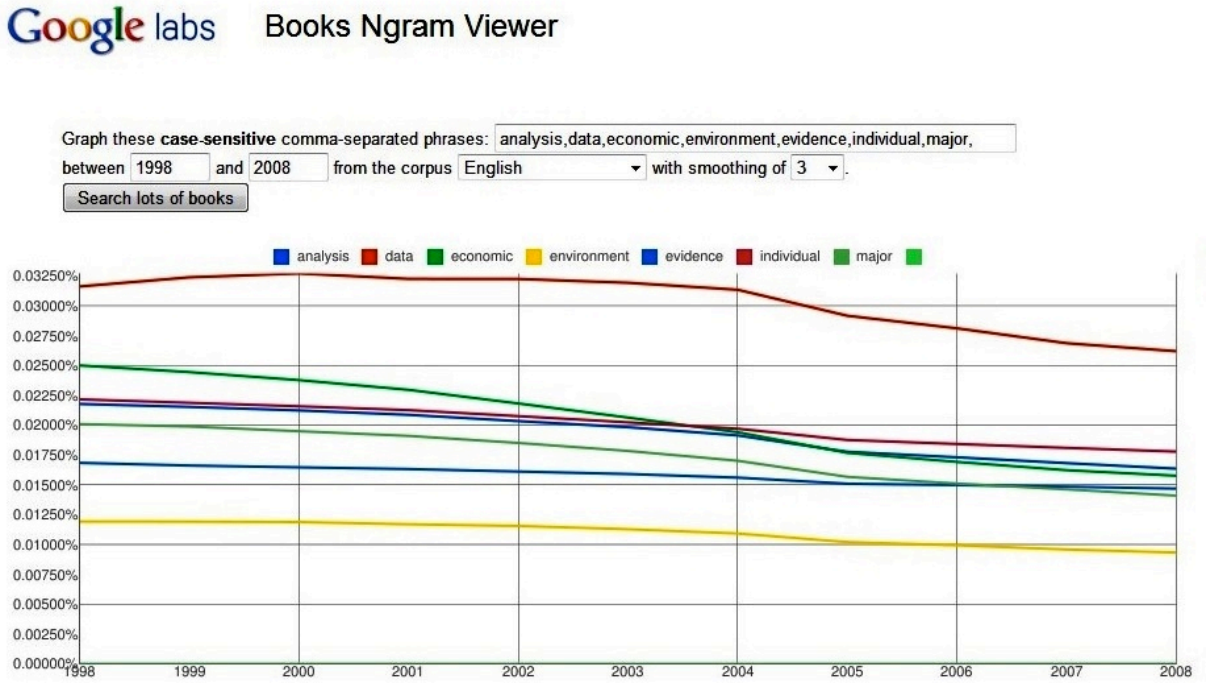

Figure 4. Frequency decline of words 1-7: *analysis, data, economic, environment, evidence, individual, major* from Google Ngram Viewer, 1998 to 2008.

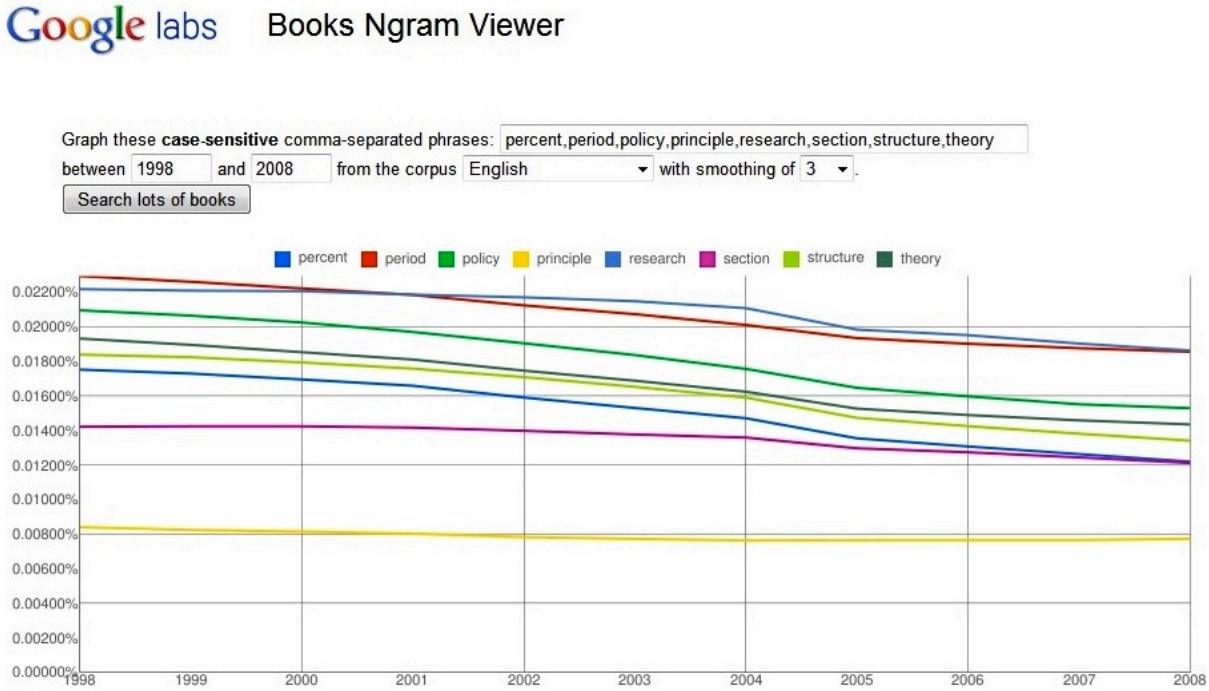

Figure 5. Frequency decline of words 8-15: *percent, period, policy, principle, research, section, structure, theory* from Google Ngram Viewer, 1998 to 2008

The graphical frequency decline of these common AWL words as illustrated by outputs produced by Google Ngram Viewer suggests that there may be a need to update AWL and similar word lists, by using mega-corpora. Academic words that are trending higher in the past 10 years should be added. A quick search for words such as *software, interface, develop*, and *document* in Google Ngram Viewer shows increasing frequencies of these words in English books. Word lists heavily rely on frequency of use, which could be more accurately provided by larger and more representative databases. While English books from Google Ngram Viewer clearly do not represent the entirety of academic writing registers, these figures give a macro-level look at frequency distributions and suggest how these could be used to gauge actual usage of words in published texts across recent time periods. This macro-level data could then be used to further track specific lemmas in a more specialized corpus of academic writing with clearly identified sub-registers such as disciplines, text types, or other genres (e.g., research report, freshmen composition, and second language writing).

4.2. Comparison of Google Ngram Viewer and COHA results

Together, the results from COHA and Google Ngram validate an overall decline of frequencies for the AWL Sublist 1 from 1998 to 2008 (from 1990 in COHA). Thirteen of the 15 words from COHA produced trending data comparable to Google Ngram Viewer's output. Discrepancies are found in the tracking numbers for two AWL words: *research* and *data*. *Research* shows a considerable increase in COHA from 164.16 tokens normalized per million to 180.91 tokens per million. This result is in contrast to what Google Ngram Viewer found for *research* as shown in the line graph (Figure 6) below. *Data* in COHA has a very slight increase from 101.00 to 101.60 tokens per million compared to Google Ngram Viewer's declining line graph.

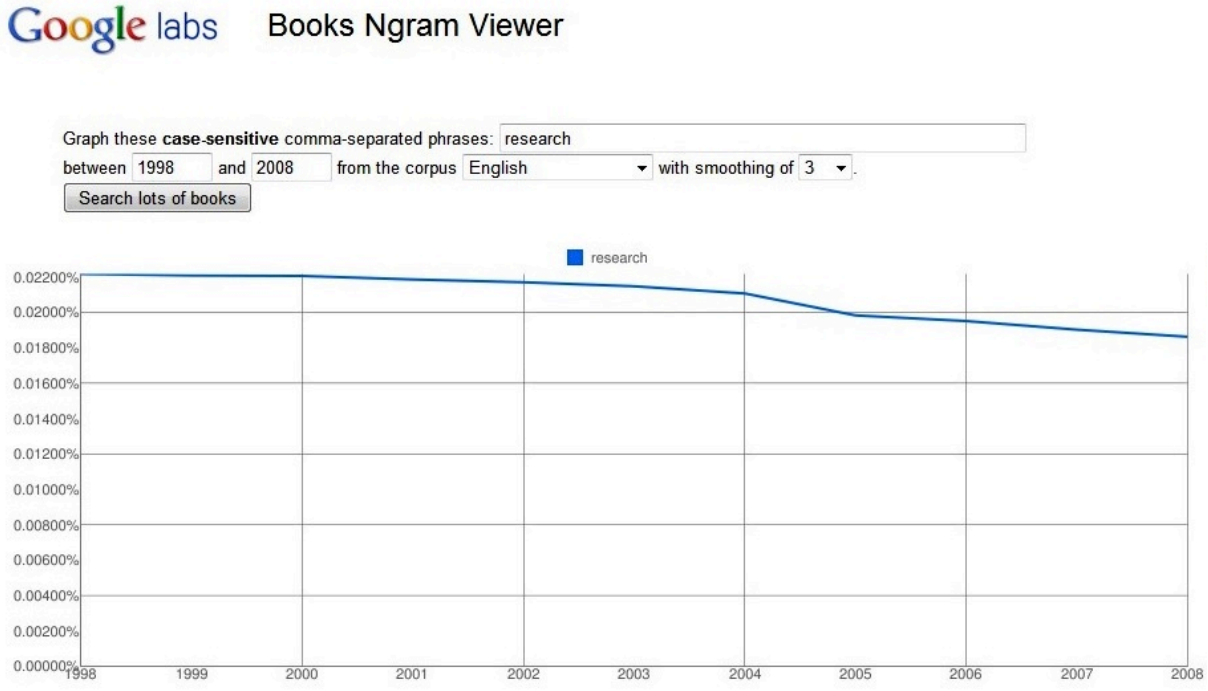

Figure 6. Frequency data for *research*, 1998 to 2008 from Google Ngram Viewer

Although COHA and Google Ngram Viewer show similar overall trends for most of the words from AWL Sublist 1, in addition to the discrepancies noted above, there are also observable differences in the micro distributions of the 15 words examined in this case study. These differences may be due to the clear distinction between the texts comprising these two mega-corpora or possibly from the major difference in total word counts. Table 2 lists words with the greatest difference in each corpus (all declining in frequency), with the words in bold indicating that the word appears on both lists and, therefore, exhibits a very similar trend.

Table 2
Comparison of micro differences between Google Ngram Viewer and COHA

	Google Ngram Viewer	COHA
1	data	**section**
2	**economic**	**economic**
3	analysis	percent
4	policy	major
5	**theory**	**theory**
6	structure	period
7	environment	evidence
8	research	individual
9	**section**	principle

Interestingly, *data*, which shows a slight increase in COHA, displays the biggest decrease in Google Ngram Viewer. *Section* has the biggest decrease in COHA - 151 tokens per million in the 1990s to 68 tokens per million in the 2000s – while it has the lowest decrease in Google Ngram Viewer. *Economic*, which demonstrates the second biggest decline in both COHA and Google Ngram Viewer, and *theory*, which has the fifth highest decrease for both mega-corpora, are the other words in common between the two.

4.3. Micro comparison of section, theory, and economic lemmas/lexemes across corpora

A closer look at the lemmas and lexemes of the three words that showed similar waning outcomes in both corpora is presented below, in order to show how word families could be tracked and compared further across time periods and possibly sub-registers. Table 3 shows the lemmas and lexemes from the AWL for each of the three words.

Table 3
Comparison of micro differences between Google Ngram Viewer and COHA

section	theory	economy
sectioned	theoretical	economic
sectioning	theoretically	economical
sections	theories	economically
	theorist	economics
	theorists	economies
		economist
		economists
		uneconomical

Results from the declining lemma, *section*, show that only *section* and *sections* are used commonly in COHA. Both *sectioning* and *sectioned* did not clearly register in the graphical outputs, although a closer look at the numbers shows that the use of *sectioned* very slightly increased during this period. Figure 7 illustrates COHA's data for these four lexemes, comparing distributions from 1990 and 2000.

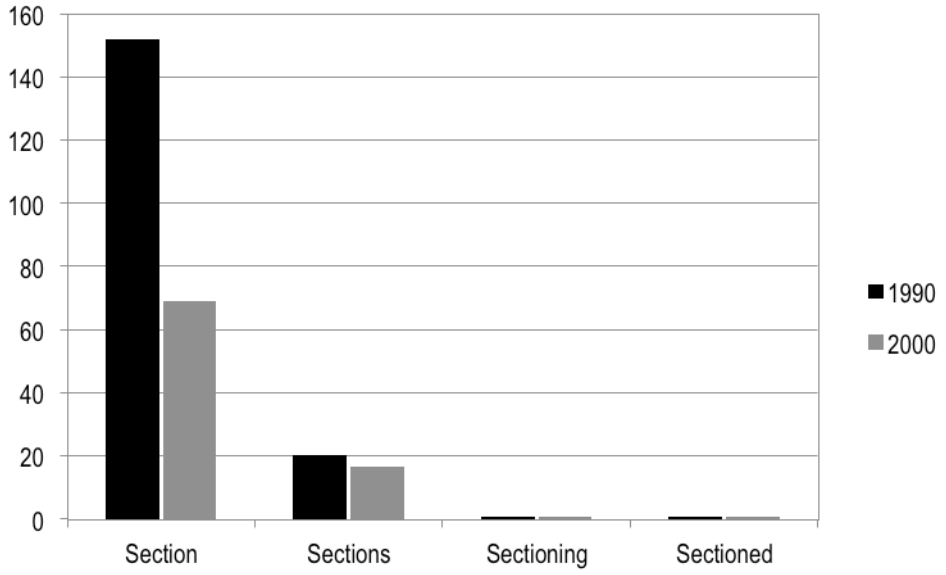

Figure 7. COHA results for *section*

The distributions of *theory* and its lexemes from Google Ngram Viewer are shown in Figure 8. *Theory* is the most frequently used word, followed by *theoretical* and *theories*. Consistent with the overall AWL findings, these words also demonstrate a general decline. *Theoretically*, *theorist*, and *theorists* exhibit very minimal frequencies of use across English books from this ten-year period.

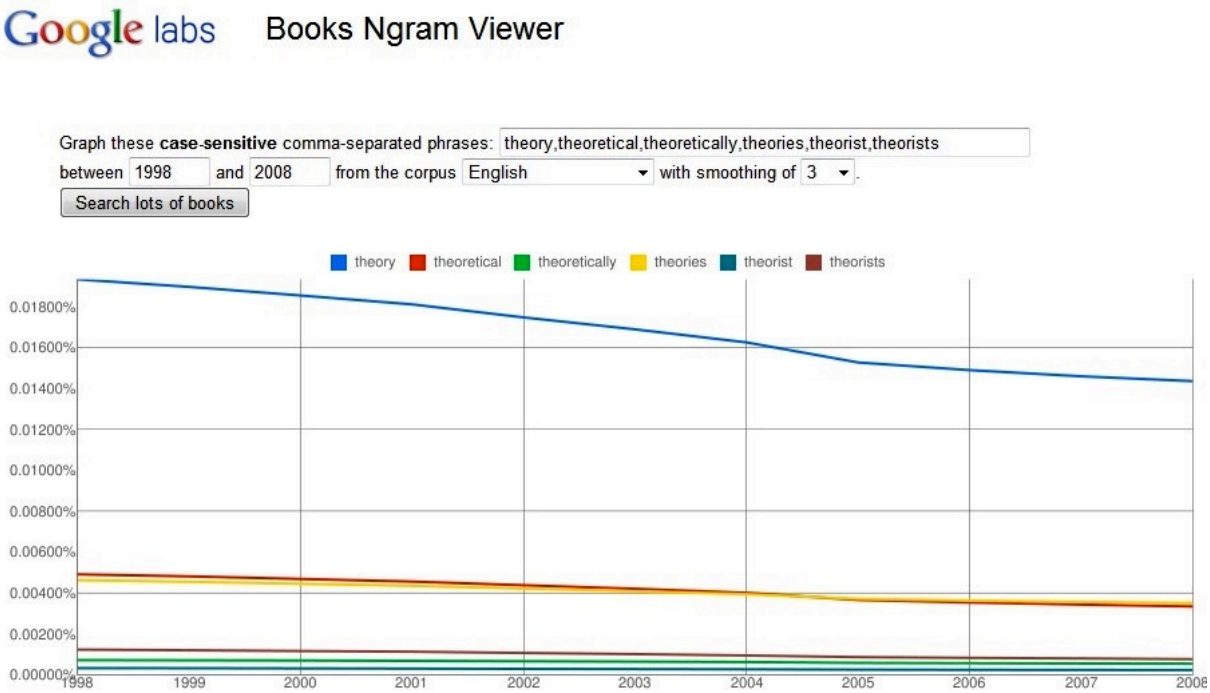

Figure 8. Google Ngram Viewer results for *theory* and its lexemes

Figure 9 from COHA displays a comparatively similar result to data from Google Ngram Viewer. Yet, one of COHA's advantages over Google Ngram Viewer in this kind of micro comparison is that COHA returns normalized tokens that show specific and easily interpretable numbers. For example, the lexeme *theorist* increased from 1.25 tokens per million to 1.79 tokens per million from 1990 to 2000, which was not

observable from Ngram's figures in Google. Overall, *theoretical, theoretically, theorist,* and *theorists* were not common in the two corpora. For the purpose of this paper - analyzing AWL - both Google Ngram Viewer and COHA show a decline in the frequently employed words *theory* and *theories*, while less common lexemes stabilize at a very low frequency or with an occasional, very slight increase in use.

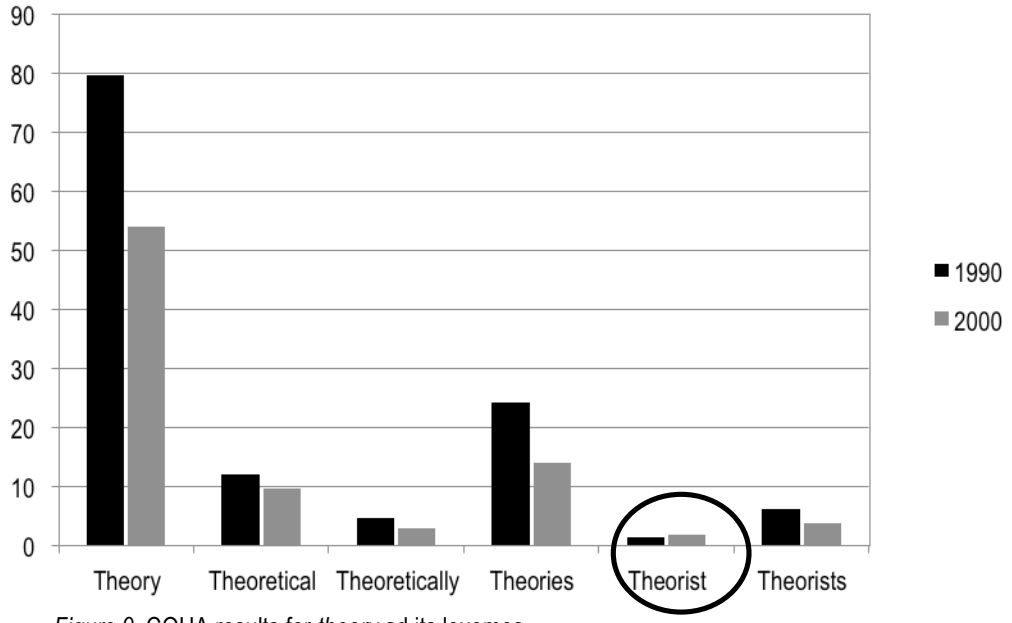

Figure 9. COHA results for *theory* ad its lexemes

Unlike *section* and *theory*, *economic* is not a lemma, but one of the lexemes of *economy*. *Economic* and *economy* are the two most commonly used words in this word family. The remaining words, *economical, economically, economics, economies, economist,* and *economists* are not as common, with less than 20 tokens per million (except for economics in 1990) in COHA (Figure 10).

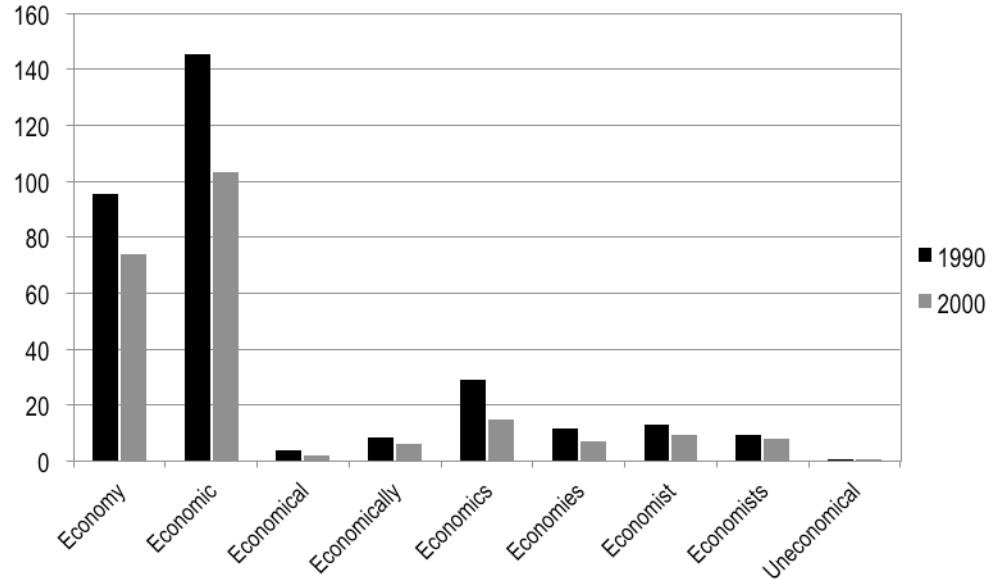

Figure 10. COHA results for *economic* and its lexemes

All lexemes in Figure 10 consistently show a decline (except for *uneconomical* which shows no change at all, but did not register bars in the figure because of very minimal normalized tokens found). In sum, Google Ngram Viewer and COHA results indicate that the lemma groups of section, *theory*, and *economic* follow very similar overall tracking trends. The most frequent lexeme is followed by another that is somewhat frequently used, both of which are in decline. These two lexemes are then followed by a variety of much less frequent words in the family that are either rarely employed or in decline. The COHA tables do show slight declines in less frequent *theory* and *economy* lexemes, while for a couple of lexemes (e.g., *theorist, uneconomical*), a slight fall or slight rise is also observed. Overall, what can be seen from these micro-comparisons explored in this case study are patterns exhibiting the consistent decline in use of the words on Sublist 1 of the Academic Word List.

5. Conclusion

This exploratory case study suggests that many of the commonly-used words listed as part of the Academic Word List (Coxhead, 2000) are declining in overall frequency of use when investigated on a macro level using mega-corpora. Both Google Ngram Viewer and COHA generally confirm this trend. Frequency data from these tools using normalized tokens and percentages support each other when presented through visual plots and estimates, and, therefore, further attest to their general consistency. It is reasonable to suggest then that Google Ngram Viewer and COHA are viable resources for macro linguistic research with AWL and similar wordlists. Further microanalysis of word distributions indicates that less-frequent lexemes have maintained a fairly stable rate of low usage. Even with very minimal tracking of normalized tokens and percentages of these less-frequent lexemes, both Google Ngram Viewer and COHA are able to produce important comparative data across a variety of words useful in projecting vocabulary shifts and changes.

These results of an overall decreasing rate of use of common AWL words are not surprising. Linguists have long noted that words have a finite life span (Crystal, 2006). Michel et al. (2011) noted that word usage is typically characterized by a spike at some point, then directly followed by a slow decline. Occasionally, words revive and have a short comeback at times, again followed by a decline measured through frequency distributions. The results from this study adhered to these general principles.

5.1. Pedagogical implications

This declining trend in vocabulary use, when measured in a relatively short timeframe (e.g., 1990 to 2008) is very relevant if applied to standard word lists used for language teaching. As the AWL has been used as a reference for teachers of academic writing for a range of learners, including non-native speakers of English, accurate distributions representing the present status of word usage in specific contexts is of great importance. Teachers and curriculum developers, then, might focus on these distributions to match how patterns of vocabulary are utilized currently outside the classroom. In fields such as English for Specific Purposes (ESP) and English for Academic Purposes (EAP), these are very important arguments for pedagogy.

Additionally, accompanying the falling frequencies of these AWL words is the consistent rise of new words. Every year, there are approximately 8,500 new words added to the English language (Michel et al., 2011). The constant influx of new words gives a writer more options in presenting ideas and structuring academic arguments or explanations. Such spread of new words certainly contributes to the variation in word lists and may cause a slow decrease in overall usage of the words similar to the ones on the AWL. Coxhead's AWL appears to be naturally heading in this direction, and there may be a need for a new list, as Coxhead herself proposed (Coxhead, 2011). At the very least, the AWL needs to be consistently updated, since it is used by many teachers for reference in genre-based instruction, especially in discipline-specific writing. With the availability of mega-corpora free to public access, there is no reason why a new, expanded word list cannot be further developed by various groups of researchers.

5.2. Comparison between Google Ngram Viewer and COHA

Google Ngram Viewer and COHA are tools that will continue to inspire the production of research studies on language variation and use. Although this study found minor discrepancies in the distributional data of some words, both corpora have shown consistently similar patterns of results, especially on a macro level, even with the obvious difference in total word counts. Statistically, when frequencies are normalized from these two databases and the investigated linguistic feature is not extremely rare, the large word count difference has very limited effect (Biber, 1993). COHA's present structure and range of features, however,

make it more ideal for corpus-based research than Google Ngram Viewer. This database is monitored and regularly updated and Davies provides users extensive "Help" features and a bibliography of COCA/COHA published studies over the years. COHA's clearly defined registers and search features such as collocates, POS-tagged data, and synonyms are grounded in linguistic theory and are well-motivated by previous research in corpus linguistics. In contrast, Google Ngram Viewer's raw database has very limited search options and interactive features at present. However, while impossible to read as individual texts, Google allows users to download n-grams which could then be processed further using specialized computer programs. COHA's database is not available for free download to users. A sampling of Google's texts of published books from 1500s to the present across major languages would represent a very important, globally available corpus for extensive linguistic analysis.

5.3. Directions for future research

Aside from an exploration of AWL word distributions in mega-corpora, one of the primary goals of this study was to contribute to the growing body of research involving the use of Google Ngram Viewer and COHA. In the past years, several diachronic language (vocabulary) change studies used these tools quite successfully, but have mostly focused on cultural and historical shifts. However, systematic, replicable research methodologies and various options in framing research questions are needed to serve as models for how these tools could be effectively applied in the fields of linguistics and corpus-based research. More validation through similar studies is necessary, including the application of statistical significance testing in examining how normalized tokens and percentages compare, beyond visual representations and line graphs. Additional comparative research with an extensive range of lexico/grammatical features and demographic information from texts will also have to be explored further.

Keuleers, Brysbaert, and New (2011) suggest that Google Ngram Viewer can be effectively utilized for many types of studies in psycholinguistics. There are also potential applications of Google Ngram Viewer for manipulation using other interfaces that aid dictionary research and linguistic tagging approaches (Sekine & Dalwani, 2010). Like Davies' (2011) attempt with the Google Books interface, other searches with clearly-defined linguistic foci could be completed in the future. The use of Google's corpora of books could be valuable in a variety of natural language processing applications. Results of comparative and contrastive analysis of linguistic data from these texts across time periods will provide fascinating information about human language.

References

Bauman, John (1995). Online. *About the General Service List*. Available at: http://jbauman.com/aboutgsl.html (accessed March 2011).

Bautista, Lourdes (2011). *Studies in Philippine English: Exploring the Philippine component of the International Corpus of English.* Manila, Philippines: De La Salle University Press.

Biber, Douglas (1990). Methodological issues regarding corpus-based analysis of linguistic variation. *Literary and Linguistic Computing, 5*(4), 257-269.

Biber, Douglas (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*(4), 243-257.

Biber, Douglas (1995). *Dimensions of register variation: A cross-linguistic perspective.* Cambridge: Cambridge University Press.

Biber, Douglas, Conrad, Susan & Reppen, Randi (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward (1999). *Longman grammar of spoken and written English.* Harlow, Essex: Pearson.

Biber, Douglas, Reppen, Randi & Friginal, Eric (2010). Research in corpus linguistics. In Robert Kaplan (Ed.), *The Oxford Handbook of Applied Linguistics*, (pp. 548-567). Oxford: Oxford University Press.

Bohannon, John (2010). Google opens books to new cultural studies. *Science*, *330*, 1600.

Bohannon, John (2011). Google books, Wikipedia, and the future of culturomics. *Science*, *331*, 135.

Cohen, Patricia (2010, December 16). New York Times online: *In 500 billion words, new window on culture*. Available at: http://www.nytimes.com/2010/12/17/books/17words.html?emc=eta1 (accessed February 2011).

Coxhead, Averil (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213-238.

Coxhead, Averil (2002). The academic wordlist: A corpus-based word list for academic purposes. In Bernhard Kettemann & Georg Marko (Eds.), *Teaching and learning by doing corpus analysis*, (pp. 73-89). Amsterdam: Rodopi.

Coxhead, Averil (2011). The Academic Word List 10 years on: Research and teaching implications. *TESOL Quarterly, 45*(2), 355-361.

Crystal, David (2006). *Language and the internet (2nd Ed.).* Cambridge: Cambridge University Press.

Davies, Mark (2005). The advantage of using relational databases for large corpora: Speed, advanced queries, and unlimited annotation. *International Journal of Corpus Linguistics, 10*(3), 307–334.

Davies, Mark (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics, 14*(2), 159–190.

Davies, Mark (2011a). Online. *The Corpus of Historical American English and Google Books/ culturomics.* Available at: http://corpus.byu.edu/coha/compare-googleBooks.asp (accessed August 2011).

Davies, Mark (2011b). Online. *N-grams and word frequency data from the Corpus of Historical American English (COHA).* Available at: http://www.ngrams.info (accessed December 2011).

Davies, Mark (2011c). Online: *Google Books corpus*. Available at: http://googlebooks.byu.edu/ (accessed December 2011).

Friginal, Eric (2009). *The language of outsourced call centers*. Philadelphia: John Benjamins Publishing.

Gao, Jianfeng, Nguyen, Patrick, Li, Xiaolong, Thrasher, Chris, Li, Mu & Wang, Kuansan (2010). A comparative study of Bing web n-gram language models for web search and natural language processing. *ACM SIGIR Forum, 44*(2), 59-64.

Grieve, Jack, Biber, Douglas, Friginal, Eric & Nekrasova, Tatiana (2010). Variation among blogs: A multi-dimensional analysis. In Alexander Mehler, Serge Sharoff & Marina Santini (Eds.), *Genres on the web: Corpus studies and computational models,* (pp. 45-71). New York: Springer-Verlag.

Herring, Susan & Paolillo, John (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics, 10*(4), 439-459.

Hilpert, Martin (2011, April). *Visualizing language change with diachronic corpus data: Introducing the flipbook technique*. Paper presented at the Exploring the Boundaries and Applications of Corpus Linguistics Conference, Tuscaloosa, AL.

Hotz, Robert Lee (2010, December 17). Wall Street Journal online: *Word-Wide Web launches new Google database puts centuries of cultural trends in reach of linguists*. Available at: http://online.wsj.com/article/SB10001424052748704073804576023741849922006.html?mod=WSJ_article_related (accessed March 2011)

Keuleers, Emmanuel, Brysbaert, Mark & New, Boris (2011). An evaluation of the Google Books ngrams for psycholinguistic research. In Kay-Michael Würzner & Edmund Pohl (Eds.), *Lexical resources in psycholinguistic research volume 3,* (pp. 23-27).Potsdam, Germany: Universitätsverlag Potsdam.

McEnery, Tony, Xiao, Richard & Tono, Yukio (2006). *Corpus-based language studies: An advanced resource book.* New York: Routledge.

Michel, John-Baptiste, Shen, Yuan Kui, Aiden, Aviva, Veres, Adrian & Gray, Matthew (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*, 176-182.

Miller, Donald (2011). *Using internal criteria to assess corpus representativeness for lexical studies.* Paper presented at the 2011 American Association for Corpus Linguistics Conference, Atlanta, GA.

Ming-Tzu, Kim & Nation, Paul (2004). Word meaning in academic English: Homography in the Academic Word List. *Applied Linguistics, 25* (3), 291-314.

Nation, Paul & Waring, Robert (1997). 'Vocabulary size, text coverage, and word lists'. In

Norbert Schmitt & Michael McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy,* p. 6-19. Cambridge: Cambridge University Press.

Nelson, Gerard (1996). 'The design of the corpus'. In Sidney Greenbaum (Ed.), *Comparing English worldwide: The International Corpus of English,* p. 27-35*.* Oxford: Clarendon Press.

Preiss, Judita; Coonce, Andrew & Baker, Brittany (2009). *HMMs, GRs, and n-grams as lexical substitution techniques – Are they portable to other languages?* Paper presented at the International Workshop: Natural Language Processing Methods and Corpora in Translation, Lexicography, and Language Learning, Borovets, Bulgaria.

Reppen, Randi (2010). *Using corpora in the language classroom.* New York: Cambridge University Press.

Sekine, Satoshi & Dalwani, Kapil (2010). Ngram search engine with patterns combining token, POS, chunk and NE information. *LREC, 2010*, 1-5.

Svartvik, John (2007). 'Corpus linguistics 25+ years on'. In Roberta Facchinetti (Ed.), *Corpus linguistics 25 years on,* p.11-26*.* Amsterdam: Rodopi.

Toor, Amar (2010). *Google Ngram Viewer gives new historical perspective on culture, language*. Available at: http://www.switched.com/2010/12/20/google-books-ngram-viewer/ (accessed March 2011).

Appendix

Complete List of 60 Words Used in the Study

Analysis	Indicate
Approach	Individual
Area	Interpretation
Assessment	Involved
Assume	Issues
Authority	Labour
Available	Legal
Benefit	Legislation
Concept	Major
Consistent	Method
Constitutional	Occur
Context	Percent
Contract	Period
Create	Policy
Data	Principle
Definition	Procedure
Derived	Process
Distribution	Required
Economic	Research
Environment	Response
Established	Role
Estimate	Section
Evidence	Sector
Export	Significant
Factors	Similar
Financial	Source
Formula	Specific
Function	Structure
Identified	Theory
Income	Variables

Eric Friginal, Georgia State University

efriginal@gsu.edu

EN **Eric Friginal** specializes in technology and language teaching, applied corpus linguistics, cross-cultural communication, and discipline-specific writing. His primary research program focuses on the exploration of professional, spoken interaction; the acquisition of fluency in ESL; and the study of language, culture, and cross-cultural communication. He makes use of corpus and computational tools as well as qualitative and quantitative research approaches in analyzing and interpreting linguistic patterning from corpora. His present work aims to contribute linguistic data that could be used for materials production and the development of training curricula in language proficiency and task performance of ESL speakers.

ES **Eric Friginal** está especializado en tecnología y enseñanza de lenguas, corpus lingüísticos aplicados, comunicación transcultural y escritura específica por disciplinas. Su principal programa de investigación se centra en el análisis de la interacción oral profesional, la adquisición de fluidez en inglés como segunda lengua y el estudio de lengua, cultura y comunicación transcultural. Utiliza corpus y herramientas informáticas, así como enfoques de investigación cualitativos y cuantitativos, en el análisis e interpretación del diseño lingüístico de diferentes corpus. Su trabajo actual pretende aportar datos lingüísticos que puedan ser utilizados para la producción de materiales didácticos y planes de estudio finalizados al dominio del idioma y la realización de tareas por hablantes de inglés como segunda lengua.

IT **Eric Friginal** si è specializzato in tecnologia e insegnamento di lingue straniere, linguistica applicata dei corpora, comunicazione interculturale e linguaggi settoriali nella comunicazione scritta. Il suo principale interesse di ricerca si concentra sull'esplorazione della lingua parlata in ambito professionale, l'acquisizione della fluidità di espressione in inglese L2, nonché lo studio di cultura e linguaggio nella comunicazione interculturale. Allo scopo di analizzare e interpretare le configurazioni linguistiche presenti nei vari corpora fa ricorso a strumenti di linguistica computazionale e dei corpora, basandosi su una metodologia di ricerca quantitativa e qualitativa. Alla base del suo operato vi è la volontà di utilizzare dati linguistici nello sviluppo di materiali e programmi di formazione mirati a una conoscenza e una competenza attive dell'inglese L2.

Marsha Walker, Georgia Institute of Technology

marsharene@gmail.com

EN **Marsha Walker** teaches various courses at the Language Institute (GaTech) focusing on the areas of student academic writing, reading, and grammar. She specializes in second language acquisition, language and literacy, classroom approaches, and corpus-based studies. Her research interests include the application of SLA theories, technology, and corpora in language teaching. She also has developed courses specific to her interests in literature and vocabulary, and she continually develops classroom teaching materials and instructional tools primarily intended for international students in U.S. universities.

ES **Marsha Walker** es docente en varios cursos centrados en escritura académica para estudiantes, lectura y gramática en el Language Institute (GaTech). Está especializada en la adquisición de segundas lenguas (ASL), lenguaje y alfabetización, enfoques didácticos y estudios basados en corpus. Entre sus intereses se incluyen la aplicación de las teorías, tecnologías y corpus de ASL en la enseñanza de lenguas. Desarrolla también materiales didácticos en relación con su interés en la literatura y el léxico, y materiales y herramientas de enseñanza específicas para estudiantes internacionales en universidades estadounidenses.

IT **Marsha Walker** insegna corsi di scrittura accademica, lettura e grammatica presso il Language Institute (GaTech). Si è specializzata nell'acquisizione di L2, nelle tecniche di lingua e alfabetizzazione, nei metodi di insegnamento in classe e nello studio dei corpora. Tra i suoi principali interessi di ricerca va ricordata l'applicazione di metodologie, tecnologie e corpus per l'acquisizione di una L2 nell'ambito dell'insegnamento di un lingua. Vasta è anche la sua produzione di materiali didattici legati al suo interesse per la letteratura e il lessico, e di materiali e strumenti educativi rivolti a studenti internazionali presenti nelle università statunitensi.

Janet Beth Randall, New York University, Tokyo

randall.janet@gmail.com

EN **Janet Beth Randall** specializes in ESL/EFL pedagogy and cross-cultural communication. Her work also consists of administration, curriculum development and pedagogically-oriented programs in English for Specific Purposes. Her current research and teaching interests include the use of corpus tools in content-based instruction and language program evaluation.

ES **Janet Beth Randall** está especializada en pedagogía de segundas lenguas, lenguas extranjeras y comunicación intercultural. Asimismo, se dedica también a la administración de programas, diseño de currículum e inglés para fines profesionales. Sus ámbitos de investigación y enseñanza incluyen el uso de corpus en programas de instrucción basados en contenidos, y la evaluación de programas de estudio.

IT **Janet Beth Randall** si è specializzata in didattica dell'inglese L2/LS e in comunicazione interculturale. Si dedica inoltre al coordinamento, alla creazione e allo sviluppo di programmi di studio e corsi di inglese per scopi specifici. I suoi ambiti di ricerca e insegnamento includono l'uso di corpus in programmi di studio basati su contenuti e la valutazione di programmi di lingua.