A3

ANALYZING AND VISUALIZING ENGLISH USING CORPORA

Various analyses of corpora can be accomplished using relatively simple computer software programs, many of which are freely available online, referred to as "freeware." In Section B3, I provide a list of these corpus tools, particularly those that are relevant to teachers, including a description of what they do, their developers, and where they can be downloaded.

The most common and most relevant corpus tool for teachers and learners is the **concordancer.** *AntConc* (Anthony, 2014), *WordSmith Tools* (Scott, 2012), and *MonoConc Pro* (Barlow, 2012) are stand-alone concordancers that are easy to use and have intuitive commands in running searches and other functions. Concordancers are also included as built-in applications in MICUSP or COCA and many other online databases. These are programs that can extract words or key words as they appear in a corpus. Word or phrase frequencies can be easily obtained, and the contexts within which these words are used can also be collected by taking words that appear before and after the designated key words in the corpus. This process is known as Key Word in Context or **KWIC**. Concordancers can also easily provide a word frequency list (from the most common word to those appearing only once), n-grams, and extract collocations of a target word or phrase. Advanced corpus researchers may need to use very specialized computer programs designed to extract particularly unique patterns that are not provided by concordancers (Friginal & Hardy, 2014a).

A3.1 Linguistic Analysis of Corpora

The following subsections provide a brief discussion of common linguistic constructs typically investigated using corpora that are useful to teachers in developing materials for the classroom. We start with basic unit-level frequency distribution,

from a single word or a phrase, then move on to KWICs, collocations, multiword units, key words, and patterns of co-occurrence of various tagged features.

lexical verbs
modal verbs
copula

A3.1.1 Frequency of Single Features

Determining the frequency of a single linguistic feature from corpora is one of the most basic types of analysis in corpus-based research. Questions such as "What are the most frequently used words in A-graded laboratory reports in Chemistry?" or "What are the top 12 most common lexical verbs in spoken American English?" are easy to extract from the relevant corpus. The former simply requires running the word list function of *AntConc*, and the latter will first require a corpus that is tagged or annotated for part-of-speech (POS), that is, the teacher will have to utilize a POS-tagger (see Section B3) to obtain the frequency of the most common lexical verbs—these are meaning-carrying, one-word verbs, such as *sing*, *talk*, *think*, or *find* and their lemmas—in the corpus. As emphasized in the previous section, frequency is important for teachers in describing the features of language varieties, including academic language, and in determining what to focus on when considering how to teach vocabulary or grammatical features. Popular word lists such as Coxhead's (2000, 2011) or Nation's (2001) "Academic Word Lists" (see Section C2) have been used in developing teaching and learning materials for students in many academic writing/speaking classes (Friginal et al., 2017). Biber (2006) noted that although most ESP/EAP studies have focused on written academic discourse, more recently, researchers have also turned their attention to university classroom discourse and the combined frequencies of various linguistic features. In addition to individual counts and frequency distributions (e.g., counts for how many pronouns or counts for '*in contrast*' or *however*) exploring the distribution of functional features, such as the study of stance and evaluation, informational discourse, and hedging in speech, has provided results for comparison across academic registers. Frequency is important in both the description of language varieties and in determining what to focus on in a vocabulary lesson. For example, it has been shown that even language specialists cannot accurately estimate the relative frequencies of words in a particular setting (Alderson, 2007). This is a paradox because many of our intuitions of existence and frequency of words, word types, and grammatical constructions are influenced by what stands out to the observer as different. Thus, casual observers of language may be more likely to perceive infrequent linguistic features as frequent (Friginal & Hardy, 2014a).

Frequency will have to be properly measured and reported. The frequency of a linguistic feature is relevant when compared with other features or when interpreted across registers. In order to make correct comparisons and interpretations of frequency data, **normalized frequency (nf)** will have to be presented. Relative frequency can be determined by calculating the frequency

of the construct per *x* number of words. Depending on the feature and the size of the corpus, a teacher might choose to measure the number of occurrences per 100, 1,000, or 1,000,000 words. This process is also called normalizing (i.e., normed count or normed frequency). In many of my studies of word/grammatical constructions, I normalize the number of instances per 1,000 words, following a simple calculation:

$$nf = \frac{\text{number of occurrences}}{\text{total number of words}} \times 1{,}000$$

Normalization not only allows for teachers to compare linguistic features with one another but also, more importantly, allows us to compare texts and corpora of differing lengths.

So, returning to my earlier question about the top 12 most common lexical verbs in spoken American English, normalized frequency data is actually available for determining this. Figure A3.1 shows the top 12 lexical verbs obtained from the Longman corpus (Biber, 2004).

Biber reports that these 12 verbs are very common in spoken interaction, and they alone will comprise close to 50% of instances of lexical verb use in the corpus. Based on these frequencies, teachers may start a lesson on teaching verbs in conversation by focusing on introducing the forms and functions of the first five: *get*, *go*, *say*, *know*, and *think*. University IEP students who are in their first semester in the US in an English oral communication class may directly benefit from this activity as they will hear these common verbs very

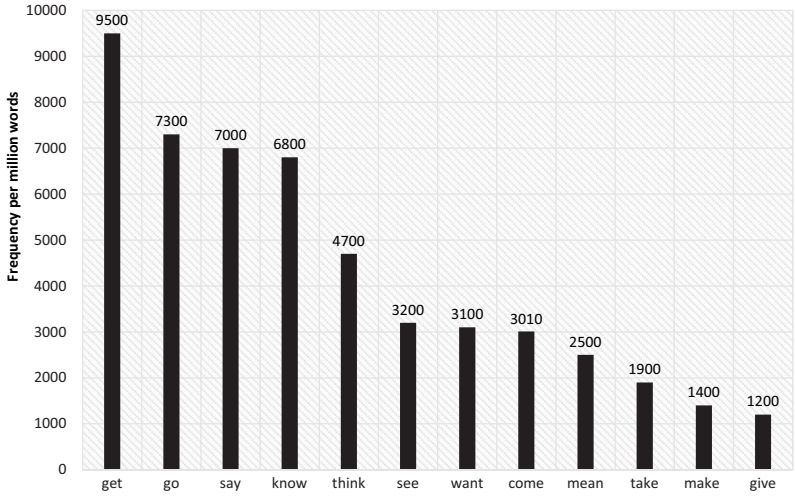

FIGURE A3.1 Top 12 most common lexical verbs in spoken American English, normalized per one million words. Adapted from Biber (2004).

frequently in interactions in and outside of the university setting. The following text extracts illustrate the various forms and meanings of *get* in conversation or informal speech:

TEXT SAMPLES A3.1 Forms and meanings of *get* in conversation

Obtaining something (activity): Check if they can **get** some of that bread.
Moving to or away from something (activity): **Get** in the car.
Causing something to move (causative): We ought to **get** these wedding pictures into an album.
Causing something to happen (causative): Uh, I *got* to **get** Max to sign one, too.
Changing from one state to another (occurrence): So I'm **getting** that way now.
Understanding something (mental): Do you **get** it?

A3.1.2 Concordances and KWICs

Traditionally, concordances are reference books comprised of alphabetical listings of all significant content words in the source material, excluding grammatical and functional words (e.g., prepositions, articles, adverbial phrases). In addition to this alphabetized index of primary words from the source text, a secondary list of words that co-occur before or after the primary word elsewhere in the text is also provided, which enables users to understand the contextual meanings of each word in the material. Scholars of the Bible, the Qur'an, and other significant religious and historical texts created concordances for these documents manually before computers expedited the task. Concordances are provided today in study or teaching editions of the Bible as appendices or footnotes, and early editions of literary works by Shakespeare, Socrates, and Homer, for example, have concordances that facilitate cross-referencing of relevant words, terms, and repeated word usage. These concordances are useful in helping identify key words and, very importantly, in defining the subtle nuances and semantic meanings intended by authors in the various, particular contexts that are essential to a complete understanding of the texts. Concordances often provide additional author commentaries, biographer footnotes, and editor narratives (Friginal, 2015).

Concordances derived from digital text files of actual language usage by speakers and writers in particular groups can provide comparative qualitative and quantitative data useful in characterizing the shared meanings of those in the defined group. Concordances can be utilized to identify the different usages and frequency of a content word, examine word collocations, explore the distribution of key terms and phrases, and create a list of multiword units. All of these additional features can be produced immediately from *AntConc*, and the resulting concordance lines can be saved for additional qualitative coding and analyses. Cross-comparisons of these concordances and their distributions across groups of speakers/writers may be invaluable in

applied linguistics. Text Samples A3.2 show KWIC lines for the phrase *in my opinion* from a corpus of personal blogs written by women based in the US (collected by Samford, 2013).

TEXT SAMPLES A3.2 Concordance lines for *in my opinion* in personal blogs

1	stumbling expression	(**in my opinion** anyway).	I mean when I try writing
2	They are not good drivers	**in my opinion**.	And what sucks is teens
3	it was a really good movie	**in my opinion**.	But it brought me to tears
4	Things change! And	**in my opinion**	they still make great music.
5	Weekends are catch up days	**in my opinion**.	You get two whole days in a week
6	a good person to be a nurse	**in my opinion**.	I'm not mean to many people
7	time to spend with someone	**in my opinion**.	We accept each other for who we
8	cause she deserves it	**in my opinion**.	look i have a nasty side too.
9	It's not very exciting	**in my opinion**.	Jazz isn't something I'd of picked
10	(for $559, 000- which	**in my opinion**	is NEVER going to sell).
11	brought up whatsoever.	**In my opinion**	the fact that we have gone 12 years
12	I have ever heard	(**in my opinion**).	His life is nothing short of a miracle

A3.1.3 Collocations

As noted earlier, the way in which linguists regard and examine discrete linguistic elements, such as words and phrases, has been strongly influenced by the work of Firth (1957). These elements should not be regarded or treated as independent from rules and other words in a text. Accordingly, the corpus approach allows for the determination of statistically significant word combinations, that is, word collocations, in a given text and how these combinations are distributed across registers. Collocations can also be found using more objective measurements from statistical results obtained from reference corpora. Prediction models of what might follow or precede a word, a noun, or a verb can be measured based on their expected frequencies. Table A3.1 shows the collocates changing over time from older to more recent for *women*, *art*, *fast*, *music*, and *food* (Davies, 2017a).

TABLE A3.1 Google Books' (from the BYU collection) changing collocates over time for *women*, *art*, *fast*, *music*, and *food* (Davies, 2017a)

	Older period	*More recent period*
women	1930–1950s: *ridiculous*, *plump*, *loveliest*, *restless*, *agreeable*	1960–1980s: *battered*, *militant*, *college-educated*, *liberated*
art	1830–1910s: *noble*, *classic*, *Grecian*	1960–2000s: *abstract*, *Asian*, *African*, *commercial*
fast	1850–1910s: *mail*, *train*, *horses*, *steamers*	1960–2000s: *food*, *track*, *lane*, *buck*
music	1850–1910s: *delightful*, *exquisite*, *sweeter*, *tender*	1970–2000s: *Western*, *black*, *electronic*, *recorded*
food	1850–1910s: *spiritual*, *insufficient*, *unwholesome*, *mental*	1970–2000s: *fast*, *Chinese*, *Mexican*, *organic*

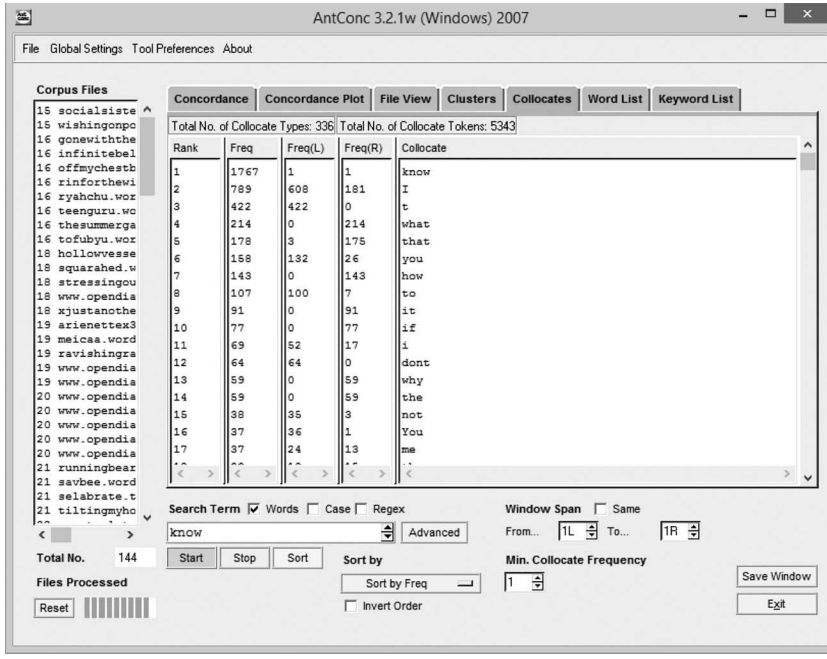

FIGURE A3.2 *AntConc's* (Anthony, 2014) first left and first right collocations for the word *know* from a blog corpus.

AntConc's first left and first right collocations for the word *know* are provided in Figure A3.2 from the same corpus, with 584,714 words, of personal blogs referenced earlier, written by women bloggers (Samford, 2013). The most frequently occurring left collocate of *know* is "I" (*I know*, occurring 608 tomes), while the most frequent right collocate is "what" (*know what*, occurring 214 times). A

contraction (*'t*), often from *don't know*, appeared 422 times in the corpus. In interpreting the *AntConc* output, disregard the search word that is listed as Rank 1 (*know*) and focus on the raw frequency reported in the output window. Users can download the full result saved as a text (.txt) file. The procedure for running collocations in *AntConc* is pretty straightforward:

1 Load the corpus: File—Open File(s)—then select your folder where your text files are located
2 Select the tab option for "Collocates" at the top of the main results window (between "Clusters" and "Word List")
3 Type your search term (*know*) in the search bar
4 Identify your first left or first right options and minimum collocate frequency (below "Window Span")
5 Click "Start" and results (Figure A3.2) will be produced.

FOR THE TEACHER

Interpreting collocations

An online article by "vaughanbell" (2017) published by Mind Hacks (https://mindhacks.com/), a neuroscience and psychology news and opinion site, notes that there is a preference for mental health practitioners to avoid the phrase *commit suicide*. These practitioners argue that *commit* refers to a **crime,** and this increases the stigma against what should be regarded as an act of desperation that deserves compassion as opposed to condemnation. The author added the following supporting arguments:

- *The Samaritans'* media guidelines discourage using the phrase *commit suicide*: Avoid labeling a death as someone having *committed suicide*. The word *commit* in the context of suicide is factually incorrect because it is no longer illegal.
- The Australian Psychological Society's *InPsych* magazine recommended against using the phrase because the word *commit* signifies not only a crime but a religious sin.

On the surface level, vaughanbell argues, claims that the word *commit* necessarily indicates a crime are clearly wrong. We can *commit money* or *commit errors*, or commit ourselves to work harder, for instance, and no crime is implied.

After examining traditional dictionary definitions of *commit* (e.g., from Google's default dictionary: [*commit*] carry out or perpetrate [a mistake,

crime, or immoral act]), vaughanbell used COCA's collocation analysis to gather the following results. I provide the first 20 collocates of *commit* in contemporary American English with their relative frequency:

(1) *Suicide*	1151	(11) *Himself*	73
(2) *Crimes*	314	(12) *Adultery*	68
(3) *Themselves*	251	(13) *Yourself*	66
(4) *Murder*	227	(14) *Acts*	63
(5) *Such*	120	(15) *Myself*	51
(6) *Ourselves*	100	(16) *Fraud*	50
(7) *Itself*	86	(17) *Crime*	39
(8) *Any*	80	(18) *Atrocities*	38
(9) *Perjury*	79	(19) *Genocide*	34
(10) *Violent*	74	(20) *Troops*	32

I have used an activity like this many times in my classes to allow students to reflect and share thoughts on an issue and then comment on what corpus data provide. It is certainly encouraging to witness popular culture's acknowledgment of corpus approaches in analyzing professional discourse. In small groups, discussion guide questions such as the following could be provided after students have read a short article. In my experience, these questions always encourage active participation and immediate use of the COCA database, with students using their phones or laptops to access the site:

1 What are your initial comments/impressions (first thing that came to mind or jumped out) after reading this article and exploring collocational data? Please share your thoughts with your group.
2 What other combinations [_____ + suicide] are possible? The author should also have considered searching for SUICIDE collocations in COCA. You can do this on your own.
3 With the list of the common collocates of *commit* shown earlier, do you think mental health practitioners who are discouraging the use of the phrase *commit suicide* are justified?

一人目ここまで

A3.1.4 Key Word Analysis

A **key word analysis** identifies significant differences in the distribution of words used by speakers or writers from two corpora. Scott (1997) defines a key word as "a word which occurs with unusual frequency in a given text" (p. 236).

This "unusual frequency" is also referred to as the *keyness value* of this word and is based on the likelihood of occurrence of the word in a target corpus as determined by a process called cross-tabulation. In other words, keyness draws from word frequency data, but instead of simple averages, statistical computation is used to determine if a word is more or less likely to occur in one corpus vs. another.

Key word comparisons provide an interesting look at the unique features of one type of discourse, language variety, or register compared to another. Key words can be extracted easily using *AntConc* and *WordSmith Tools*. Note that this process involves loading a target corpus, also known as "node corpus," and a reference corpus into the software to proceed with the analysis. A video tutorial for running key word analysis using *AntConc* is available from YouTube. Search: "*AntConc* – Keywords."

In the following example (Table A3.2), I provide two key word lists from a collection of essays written by L2 university students responding to two argumentative email prompts. The focus here is to investigate topic effect and whether a certain topic may have an influence in writing quality. For this key word analysis, I wanted to categorize the distribution of words repeated from the actual prompts. Corpus 1 are essays responding to a question about the "importance of planning for the future." Corpus 2 asks about the implication of too much "emphasis on personal appearance and fashion." Frequency and keyness values are provided for each key word.

FOR THE TEACHER

Students in a CL class can be asked to interpret the data from the table (Table A3.2). It's a good idea to provide the additional key words, if possible, the first 100 per corpus. Clearly, students identify words specifically mentioned in the prompts as they write their responses, and these were the primary key words per corpus. First person pronoun *I* was the top key word in the "appearance" corpus. The misspelled words "apperance" and "fashons," misspelled 75 and 66 times, respectively, are both in the top 30 for Corpus 2. Teachers can ask students the following questions after they analyze the results:

1 What patterns did you recognize? How do you interpret the characteristics of L2 student writing from these two prompts? When compared to L1 writers, do you think there will be differences?
2 What are ideal topics of comparison for a key word analysis?
3 What are limitations in conducting key word analysis?

TABLE A3.2 Key word comparison from two groups of essays written by L2 students

Corpus 1	*Frequency*	*Keyness*	*Key word*	*Corpus 2*	*Frequency*	*Keyness*	*Key word*
Future				Appearance			
1	865	1312.833	*future*	1	666	788.214	*I*
2	781	1258.964	*we*	2	593	701.818	*appearance*
3	756	1193.287	*plan*	3	517	598.986	*fashion*
4	502	508.426	*young*	4	398	399.135	*look*
5	219	353.026	*carefully*	5	348	331.253	*personal*
6	534	313.228	*good*	6	289	321.35	*emphasis*
7	178	286.934	*planning*	7	854	288.933	*on*
8	384	269.069	*life*	8	215	254.454	*the*
9	155	249.858	*in*	9	167	197.645	*it*
10	174	234.975	*ensure*	10	155	183.443	*wear*
11	233	206.442	*still*	11	154	182.26	*in*
12	127	204.723	*it*	12	160	178.816	*dress*
13	110	177.319	*the*	13	158	176.474	*clothes*
14	608	172.87	*we*	14	215	175.691	*put*
15	99	159.587	*however*	15	146	172.792	*clothing*
16	844	155.405	*you*	16	143	169.241	*this*
17	182	153.326	*while*	17	141	166.874	*people*
18	94	151.527	*for*	18	119	140.837	*wearing*
19	94	151.527	*plans*	19	114	134.92	*having*
20	86	138.631	*if*	20	221	126.197	*society*
21	74	119.287	*so*	21	108	118.057	*media*
22	84	118.776	*goal*	22	93	110.066	*they*
23	73	117.675	*when*	23	225	97.515	*too*
24	93	116.337	*early*	24	82	97.047	*for*
25	190	115.973	*he*	25	313	90.27	*much*
26	386	108.578	*will*	26	75	88.763	***apperance***
27	312	105.719	*your*	27	71	84.029	*women*
28	175	102.906	*best*	28	70	82.845	*when*
29	299	102.498	*my*	29	66	78.111	***fashons***
30	79	100.017	*career*	30	66	78.111	*there*

A3.1.5 Multiword Units (MWUs) and Prefabricated Chunks

As with collocations, some words frequently co-occur as linear, formulaic strings, like a prefabricated 'chunk' of language. MWUs include a range of studies of extended strings of language, and there are various ways and operationalizations (including definition of terms) to explore this construct of formulaic language using corpus tools. Three of the commonly used approaches to MWUs are n-grams, lexical bundles, and p-frames.

N-grams. The most basic construct associated with MWUs is that of the **n-gram**. The n stands for any number variable (e.g., 4-gram = *on the other*

TABLE A3.3 The 50 most common 4-grams from the Enron Email Corpus

	Frequency	*4-gram*		*Frequency*	*4-gram*
1	87	*you have any questions*	26	19	*a copy of the*
2	82	*me know if you*	27	19	*I look forward to*
3	77	*Let me know if*	28	19	*will let you know*
4	73	*I would like to*	29	19	*you get a chance*
5	70	*Please let me know*	30	17	*I m going to*
6	67	*if you have any*	31	17	*I will not be*
7	60	*let me know if*	32	17	*please let me know*
8	44	*know if you have*	33	16	*be out of the*
9	39	*I don t know*	34	16	*I don t have*
10	38	*If you have any*	35	16	*I will be in*
11	31	*Let me know what*	36	16	*let me know what*
12	28	*I m not sure*	37	16	*ll let you know*
13	27	*give me a call*	38	16	*when you get a*
14	26	*have any questions or*	39	15	*don t know if*
15	26	*I will be out*	40	15	*Give me a call*
16	25	*out of the office*	41	15	*I will let you*
17	24	*I don t think*	42	15	*me know if I*
18	23	*Thanks for your help*	43	15	*Thank you for your*
19	23	*You have two cows*	44	15	*to be able to*
20	22	*and let me know*	45	15	*to let you know*
21	22	*I am going to*	46	15	*will be able to*
22	22	*me know what you*	47	14	*I just wanted to*
23	22	*will be out of*	48	14	*if you need anything*
24	21	*know if you need*	49	14	*me know when you*
25	21	*Talk to you soon*	50	14	*not be able to*

hand). N-grams can also be extracted using most basic corpus packages; both *AntConc* and *WordSmith Tools* have intuitive commands for n-gram extraction. Table A3.3 shows a list of the 50 most common 4-grams from a corpus of professional, workplace emails from the Enron Email Corpus (see also Section B1).

Lexical bundles. One particular type of n-gram is the *lexical bundle*, an n-gram with additional specifications as to how they are extracted or categorized. Customarily, lexical bundles consist of at least three words (tri-grams) that occur frequently—frequency determined by the researcher—across a corpus of at least one million words. Another important criterion for labeling MWUs as lexical bundles is that they must appear in at least five different texts in the corpus, that is, they are common in other registers as well. This is necessary to avoid any idiosyncratic language usages (Cortes, 2004).

P-frames. Researchers have also moved beyond looking only at contiguous strings of words to also examine frequent, patterned constructions. P-frames are consistent phraseological structures that allow, however, for variability in one position of the phrase frame. An example of a p-frame, found by Römer (2010), is *it would be* * *to*, in which the asterisk represents an open slot. Grammatically,

any number of adjectives might go into the blank slot in this example. Römer found that the most frequently occurring words in a corpus of student essays in the "blank" slot were *interesting*, *useful*, *nice*, and *better*, these accounting for 77% of all the variants in the corpus.

A3.1.6 Vocabulary Usage and Lexico-Syntactic Measures: Cohesion, Complexity, Sophistication, and Others

It has been well-documented that vocabulary development in spoken and written discourse is critical in both the literacy development and academic success of L2 learners. More specifically, students' academic success depends upon their developing the specialized and sophisticated vocabulary of academic discourse that is distinct from conversational language (Francis et al., 2006). Corpus tools may be utilized to extract and then interpret the nature of vocabulary usage by learners across levels of proficiency. For example, a number of studies have identified particular linguistic features (e.g., subordination, prepositions, linking adverbials, etc.) that are predictive of scores given by instructors/raters as well as features that distinguish differences among various academic disciplines (Römer & Wulff, 2010) and demographic factors: for example, language proficiency levels and graduate vs. undergraduate (Grant & Ginther, 2000; Hinkel, 2002).

Identifying features indicative of quality speech and writing—especially those that are discipline-specific—is of obvious pedagogical importance to teachers. An understanding and description of linguistic complexity is important insofar as it may relate to the amount of discourse produced by learners as well as the quality of that discourse, including the types and variety of grammatical structures; the organization and cohesion of ideas; and, at the higher levels of language proficiency, the use of text structures in specific genres. These features may be defined and operationalized in the development of teaching materials for the classroom. Measures such as t-units, clause constructions, type/token ratio, and markers of information density and elaboration have all informed the creation of lessons and test prompts in the L2 classroom, especially in the university setting (Friginal et al., 2017).

Computational tools, such as *Coh-Metrix* (see, e.g., Crossley & McNamara, 2009) and those developed by Scott Crossley (Georgia State University) and Kris Kyle (University of Hawaii) and their colleagues entitled "Suite of Linguistic Analysis Tools" or SALAT (http://www.kristopherkyle.com/), can be used to rate the readability and also to extract frequency counts for a range of linguistic features (see additional discussion about *Coh-Metrix* and SALAT in Section B3). These tools are more applicable for researchers and teachers than for language learners themselves. Teachers may find them useful in materials development for such topics as distinguishing between authors of texts; distinguishing between writers' country of origin, for example L2 writers from the International Corpus of Learner English (ICLE); identifying changes in

L2 language over time; distinguishing between L1 and L2 writers; classifying spoken and written registers; distinguishing parts of a paper (abstract, introduction, methods); and many others.

A3.1.7 Patterns of Linguistic Co-Occurrence

The concept of linguistic co-occurrence suggests that the linguistic composition of a particular language or discourse domain, such as face-to-face classroom interaction or a study group, may have higher frequencies of questions and responses, inserts, dysfluent markers (e.g., filled pauses—*uh*, *um*), and back channels (e.g., *uh-huh*) than that of speakers in other settings. At the same time, any given feature may not be as common in different settings such as extended and prepared lectures, news reports, or formal speech. Linguistic features, such as pronouns, past tense verbs, and nouns, often occur together whenever speakers engage in everyday conversations or talk about their previous experiences and recent events. These same features might also appear frequently in written, first-person narratives or soliloquies about past events. A simple KWIC search will not suffice to capture and document these co-occurring features from corpora. A more advanced statistical framework is necessary to identify the composition of features that are frequently found together within a corpus.

Corpus-based multidimensional analysis (MDA) was introduced in Biber's (1988) *Variation across Speech and Writing* as a research methodology for exploring linguistic variation in spoken and written English texts. Biber's primary research goal was to conduct a unified linguistic analysis of spoken and written registers from 23 sub-registers of the LOB (for written texts) and London-Lund Corpus (for spoken texts). He was able to substantially redefine a range of register characteristics of spoken/written discourse by using a multivariate statistical procedure to identify intrinsic linguistic co-occurrence patterns across POS-tagged texts. Subsequently, he was able to establish a model of corpus-based research that could be applied to even more specialized contexts. MDA relies on factor analysis (FA), which identifies the sequential, partial, and observed correlations of a wide-range of variables in order resulting in groups of co-occurring factors (Friginal & Hardy, 2014b).

Biber's Factor 1, interpreted as *Involved vs. Informational Production*, is characterized by the combination of private verbs (e.g., *think*, *feel*), demonstrative pronouns, first- and second-person pronouns, and adverbial qualifiers as speakers or writers talk about their personal ideas, share opinions, and involve an audience with the use of *you* or *your.* This discourse is also informal and hedged (*that* deletions, contractions, *almost*, *maybe*). The contrasting features include the giving of information ("Informational Production") as a priority in the discourse. There are many nouns and nominalizations (e.g., *education*, *development*, *communication*), prepositions, and attributive adjectives (e.g., *smart*, *effective*, *pretty*) appearing together with very few personal pronouns. This suggests that the focus is upon informational data and descriptions of topics rather than upon

the speaker or writer. Additional characteristics of this production are more unique and longer words (higher type/token ratio and average word length) and greater formality in structure and focus.

FOR THE TEACHER

Using Biber's MDA approach, Hardy and Römer (2013) extracted dimensions of A-graded university writing from MICUSP. Their Dimension 1 distinguished between *Involved, Academic Narrative*, very common in Philosophy and Education papers, and *Descriptive, Informational Discourse*, typical of A-graded papers in biology and physics. The following text samples show a biology report compared to a philosophy critique. What characteristics are typical of one text sample in contrast to the other? What useful teaching applications occur to you as you identify such patterns? See a brief additional discussion on the application of MDA results to pedagogy in Section B1 from the MICUSP description.

TEXT SAMPLES A3.3 Comparison of involved, academic narrative and descriptive, informational discourse in MICUSP

BIO.G0.25.1, report, final year UG, NS

Normally malaria is a curable disease, but only if treated properly. After an infectious bite there is an incubation period in the host that varies depending on the species of Plasmodium, before there is an onset of symptoms. The symptoms of malaria that a human host will go through can be categorized as either uncomplicated or severe. With uncomplicated malaria, the symptoms last between 6–10 hours and include a cold stage, a hot stage and then finally a sweating stage. Symptoms occur in a mixture of fever, chills, sweats, headaches, nausea, vomiting, body aches, and general malaise.

PHI.G0.06.1, critique/evaluation, final year undergrad (UG), Native speaker (NS)

Socrates then concludes that group (D) does not exist, since those people, by desiring what they believe to be harmful (bad) things are desiring to be miserable and unhappy. No one wants to be miserable and unhappy, so no one desires what he believes to be bad. (A)–(C) actually desire what they believe to be good, and group (D) does not exist, so no one desires what he believes to be bad. I feel compelled to say here that although Socrates actually claims that "no one wants what is bad" (78), what he means is that no one wants what he believes is bad.

A3.2 CL and Visualization of Linguistic Data

Various methods of visualizing linguistic data have resulted from the relatively easy processing of corpus-based frequencies and transforming of these data into figures or images. From simple bar graphs or histograms to more complex, online interactive semantic maps, CL approaches have produced excellent visual representations of language and innovative approaches to their use in the classroom. Technically, concordancers are also visualizers, able to highlight KWICs as they appear in the corpus. Visualizers are important in language learning because they add another layer of information that is not fully captured by texts (i.e., letters and numbers) alone. They break the monotony of the written page, provide teachers a creative output in sharing data, and also effectively address the needs of visual learners.

CL-based frequencies have been used in **infographics** (i.e., information + graphics), which are now very common in online articles and advertisements. These are visual representations of data or knowledge intended to present information quickly and clearly. Smiciklas (2012) notes that infographics can develop reader cognition as graphics can enhance our visual system's ability to see patterns and trends more efficiently. There are many 'drag and drop' infographic creators online like Canva (www.canva.com) or infographic makers Piktochart (https://piktochart.com/).

A3.2.1 Word Clouds and Frequency Visualizers

The most common approach for an exact-match visualization is a word cloud. A **word cloud** is a graphical representation of word frequency from a text or a corpus. The following sample word cloud (Figure A3.3), created using WordClouds.com (https://www.wordclouds.com/), represents the first 10 pages of this book, illustrating visually that the words *language*, *English*, *teaching*, *corpora*, *writing*, *tools*, *students*, *corpus-based*, *book*, and *grammar* are the most frequently repeated words. These 10 words can capture and display the overall theme of the book based on nothing other than an 'eye-balling' of what's frequently repeated in the text.

What was previously a complicated process involving computer programming, the creation of word clouds has now become an easy *cut*, *paste*, and *create* process online. Word cloud generators convert frequency data into a graphical outline of text content. 'Tags' are identified from single words, and the importance of each tag, defined as frequency of appearance in the text, is shown with increased font size and/or change in color (Halvey & Keane, 2007). This visualized format is convenient for quickly locating the most prominent word in the input text or corpus. In Figure A3.3, I did not convert the entire text of the first 10 pages of this book into all lowercase font, including the first letters of all words, so the words *Language* and *Teaching* appear in the word cloud with

FIGURE A3.3 A word cloud of the first 10 pages of this book.

uppercase first letters. This illustrates the fact that CL extracts anything and everything that is available in the dataset, from the most frequent feature to those that only appear once (see Section C4.2 for a lesson that incorporates visualizing political speeches with word clouds). In addition to WordClouds.com, there are several other word cloud generators such as Wordle (www.wordle.net) or Tagxedo (http://www.tagxedo.com/) that provide free word or tag cloud templates and other applications. Tagxedo, for example, is also able to easily provide a key word list and the use of various color and design options.

Because CL relies on frequency data by group or by text file, it is easy to transform these distributions into figures, especially histograms and charts. From MS Excel functions to more sophisticated statistical packages like SPSS or R, figures to enhance numerical presentation are often included in CL research articles and textbooks. These figures are also easily incorporated into language classroom activities. Students in small groups or pairs can examine figures/graphs, identify patterns, and make exploratory conclusions. Figures A3.4 and A3.5 show word and tag frequency data that learners can discuss and interpret.

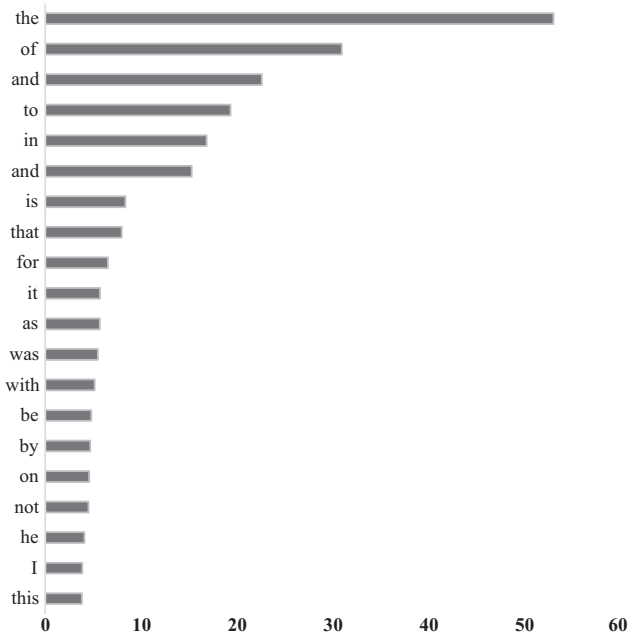

FIGURE A3.4 Visual representation of the Top 20 words in the English language from Google Books (a mega corpus of more than 500 billion words from scanned books in English and also other languages).

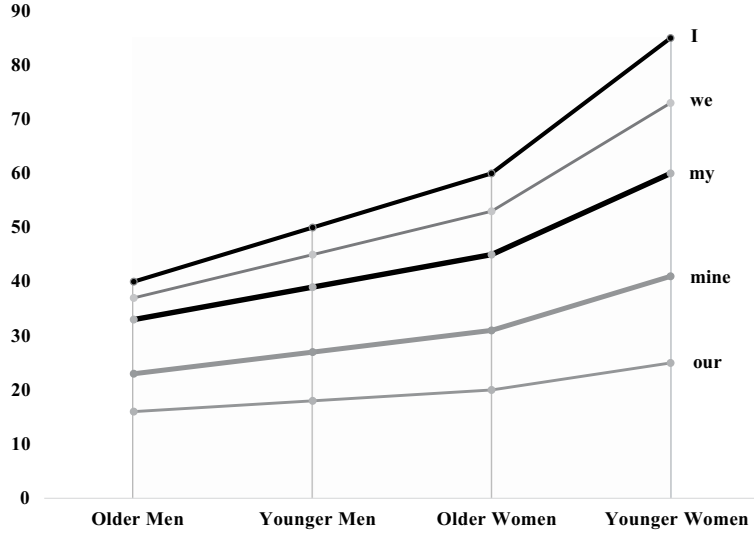

FIGURE A3.5 Use of personal pronouns *I*, *we*, *my*, *mine*, and *our* by men and women bloggers in two age groups (30 and younger vs. 31 and older). Adapted from Friginal (2009).

The comparison illustrated in Figure A3.5 shows a dramatic difference between older men and younger women in their use of personal pronouns in personal blogs, most of which were obtained from sites such as *LiveJournal* from 2006 to 2009. In the following two short excerpts, the presence of many first-person pronouns in a blog written by a 17-year-old female high school student contrasts considerably with the tone and focus of blog writing by a 67-year-old retired male. The two excerpts seem to address personal topics and were both directed to readers who were quite familiar with the bloggers.

TEXT SAMPLES A3.4 Comparison of blog texts

Oh, thank you God. Band camp really sucks. **I** am so tired of all of it! It doesn't matter, tomorrow is the last day. **I** don't really feel like updating much. Go figure. **We** have the 1st, 2nd, and up to set 15 of the 3rd song completed, but just as last year, **our** drill writer is stupid and is falling behind. **We** have no more drill to work on. Hopefully **we** will have more tomorrow. (Female blogger, high school student)

Table talk for the Sunday brunch crowd was the Senior Prom at the Golden Age Center last night. Retired biology teacher Denver Zygote and Granny Garbanzo double-dated with Judge and Mrs. Halfthrottle. The big excitement came about half-way through the festivities when Granny attempted to Watusi with her cane in her hand. (Male blogger, 60s, retired)

A3.2.1.1 Focus on Diachronic Data

Visualizing linguistic changes or trends across time is one of the primary foci of the Google Ngram Viewer (https://books.google.com/ngrams) and the Corpus of Historical American English (COHA, Davies, 2010–) also created and published by Davies. These two mega-corpora feature billions of words of American English representing various time periods. Both online visualizers present comparison data that default from the 1800s to the present, and they can extract words or any multiword combinations as they appear in the databases. For COHA, normalized frequencies of search words/phrases can be easily obtained, and the contexts within which these words are used can also be analyzed by examining KWIC lists that appear below the chart feature of the site. Figures A3.6 and A3.7 illustrate the declining usage trend of the word *gentleman* from the 1800s to the present from COHA and Google Ngram Viewer.

3人目

A3.2.2 Visualizers in the Classroom with Ying Zhu

I collaborated with Ying Zhu of the Creative Media Industries Institute and Xi He of the Department of Computer Science at Georgia State University in developing *Text X-Ray* (Zhu & Friginal, 2015), a POS-visualizer and writing platform that can be used by teachers and their students in various contexts of university-level language teaching, especially in academic writing across

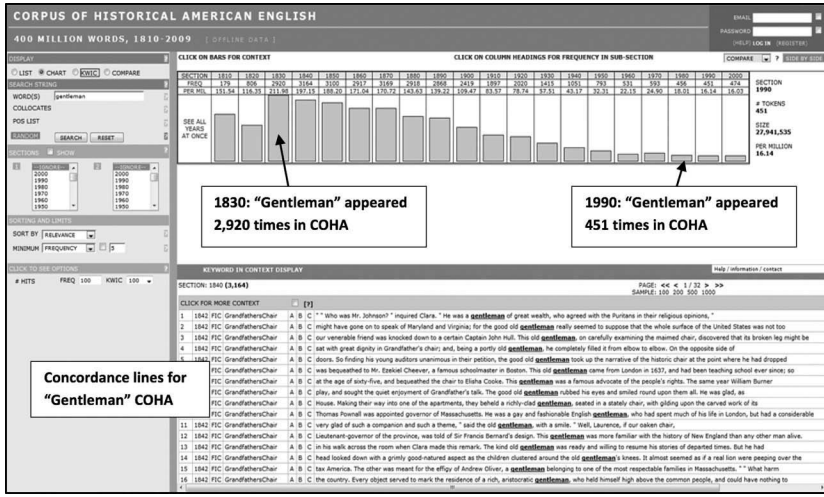

FIGURE A3.6 Distributional comparison of the word *gentleman* from the 1880s to the present in COHA, with KWIC results. Figure and illustrations adapted from Davies, 2010–.

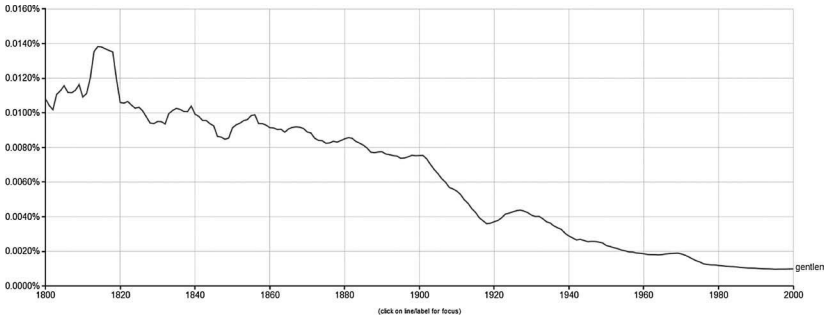

FIGURE A3.7 Frequency of *gentleman* in English books from 1800s to the present from the Google Ngram Viewer.

disciplines. Zhu leads the Hypermedia and Visualization Lab and Brains & Behavior research program at GSU, and his research interests include computer graphics, data visualization, and bioinformatics. As an L2 learner himself, and also one that identifies as a visual learner, Zhu has advocated for the use of computer-based visual data in language instruction. He believes that the structure of language, typically explored in grammar activities (e.g., tree diagrams), could be best comprehended by groups of learners when they were aided by color coded and interactive visuals. A model of networks showing nodes of sentences and the connections words have with each other allowed Zhu to fully

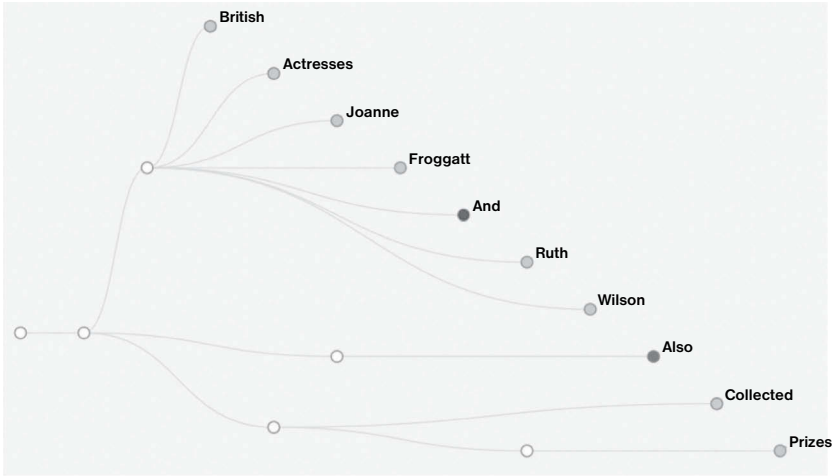

FIGURE A3.8 Sample sentence tree from Zhu. Adapted from Zhu and Friginal (2015).

appreciate the functions of various parts of speech, more so than memorizing what they mean, as required by the traditional grammar books methodology. A sentence tree program that he has developed (Figure A3.8) automatically creates sentence diagrams based on POS-tagged data.

Although lexico-grammar and writing instruction are the primary focal areas of *Text X-Ray*'s applications, the software can also be used productively for small group peer-reviewing activities in writing or peer-editing dyads that are mediated by computer technology. The design of *Text X-Ray* takes into account teachers' needs and objectives, focusing on content-based activities that can be applied to help students build academic vocabulary, learn grammatical structures, and analyze model texts, especially their own writing. Student-directed comparisons of vocabulary/POS features of texts can be facilitated through *Text X-Ray* by analyzing academic word lists and grammar patterns from teacher-prepared focal writing excerpts. Activities utilizing *Text X-Ray* may help students develop greater awareness of grammar and usage across contexts. This approach can create a great deal of positive classroom energy, encouraging students to become autonomous learners and also provide effective alternatives for students with different learning styles (i.e., student-driven learning). In Section C4.4, Berger shares sample lessons and student/instructor feedback on using *Text X-Ray* in an essay writing and editing activity.

Text X-Ray, we believe, contributes data and a range of linguistic information for learners on the structure of written texts in various academic registers. The program combines features from tools such as *Compleat Lexical Tutor* (Cobb, 2016) and *WordandPhrase.Info* (Davies, 2017b), with the addition of an interface that highlights

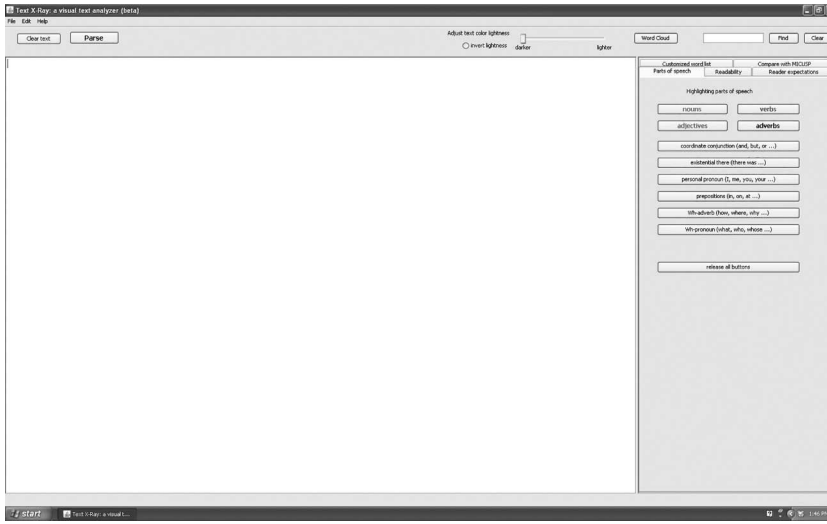

FIGURE A3.9 *Text X-Ray*'s text editor and standard application tools (Zhu & Friginal, 2015).

how POS tags are used in context. This beta version of *Text X-Ray*[1] works as a basic text editor, with built-in POS visualizer for various POS tags (e.g., nouns, verbs, prepositions) obtained from the built-in Stanford Tagger; readability and lexical diversity measures; wordlist comparisons; and a word cloud application. Another important feature of this program is its ability to compare normalized frequencies of linguistic features, for example, word/phrasal classes, with those aggregated from MICUSP. Note again that MICUSP is composed of advanced, A-graded student papers categorized primarily across disciplines and text types collected at the University of Michigan (O'Donnell & Römer, 2012). Student-produced texts can be immediately compared with MICUSP data, in real time, across disciplines, paper types, and student levels, including gender and native speaker vs. non-native speaker groups. Figure A3.9 shows the primary text editor view of *Text X-Ray* and its current set of tools and command buttons:

- File/Edit/Help—Standard application tools used to load (copy/paste) a text or obtain technical help information
- Clear Text—Button allowing users to clear/delete texts loaded on the text editor
- Parse—Command to run analysis
- Visualizer for Text Color Lightness (darker or lighter)—Color lightness control
- Word Cloud
- Search Bar (Find/Clear)
- Applications: POS, customized word list, compare with MISCUSP, readability, reader expectations

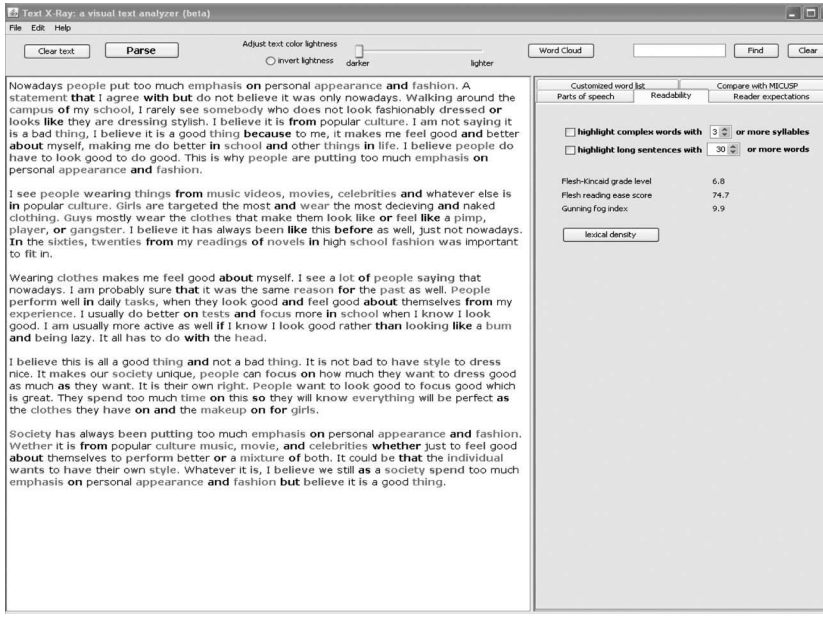

FIGURE A3.10 POS-visualizing through *Text X-Ray* (color coded POS not shown in the gray scale image, e.g., green = nouns, red = verbs, bold = prepositions).

Figure A3.10 is a sample POS-visualized essay with immediate feedback for students on options such as readability (Flesh–Kincaid Grade Level, Flesh Reading Ease Score, and Gunning Fog Index), complex words and sentences (which could be highlighted in bold), and reader expectation measures (Figure A3.10).

Structurally, the program's Visualization Engine and Visualization Interface are directly activated from a standalone browser application. *Text X-Ray* is divided into multiple panels: text panel, visualization panel, linguistic analysis panel, and control panels. The text panel displays the written input (the essay), while the visualization panel, always parallel to the text panel, enables users to analyze the texts at five levels of detail: corpus (source text), articles, words, sentences, and paragraphs. The linguistic analysis panel holds commands for readability indices, lexical density, word and sentence length, and other measures. The control panel allows users to adjust the visualization settings (colors and highlights), manage texts, and various users. Figure A3.11 illustrates the structural and programming workflow of *Text X-Ray*.

In computer-based visualization, texts should ideally be displayed alongside their visual representations. In many text visualization programs or techniques, however, visualizations often replace the texts; the original texts are often not displayed in the interface. In the context of language teaching and learning, it is necessary for the texts to be visible at all times. Visualizations should support, not replace, the original text; therefore, text visualizations will have to be linked and synchronized with the original text display (He, 2016; Zhu &

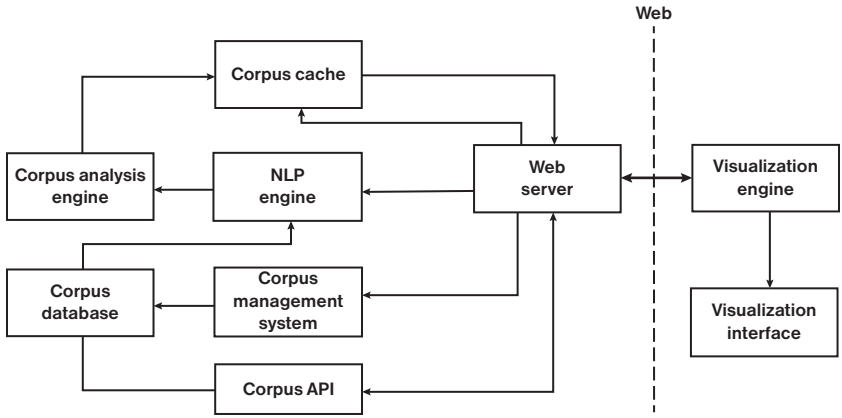

FIGURE A3.11 *Text X-Ray*'s structural workflow. Adapted from He (2016) and Zhu and Friginal (2015).

Friginal, 2015). In the classroom, data visualization will have to be introduced carefully and meaningfully. For example, because users often switch between the original texts and their visualized versions, the visualized form may have to be closer to the conventional textual display for easier mental transition and connection. Some text visualization displays are difficult for learners because they require additional mental adjustment from one form of display to another. For programming purposes, therefore, the complexity and abstract nature of innovative visualization should be controlled (He, 2016).

To investigate how *Text X-Ray* might be used in the classroom, we asked a group of users, primarily graduate students and instructors at the Department of Applied Linguistics and ESL at GSU, to pilot the software in their classrooms from 2012 to 2016. Our plan was to distribute *Text X-Ray* to a wider range of users, improve its online interface, and seek research funding to support the program for future easy access globally. The beta version of this software is still being examined for usability data from a limited release in order for us to develop and finalize the next set of improved tools that will significantly enhance the program's capabilities and usefulness. Teacher feedback was positive, overall, and as shown in the following text, there are promising applications of a program like *Text X-Ray* that teachers immediately noticed:

※このへんの教員の反応は割愛してよい

CF, Instructor, Hall County Alliance for Literacy, Gainesville, GA

> *Text X-Ray* is a program that I could sit here and play with all day because I just think it's cool that a program can pick out parts of speech in a selected text. I haven't noticed POS-tagging mistakes made by *Text X-Ray* yet, but I'm determined to stump it. My immediate thought was to introduce this program to the other instructors where I teach. Several were

interested entering their students' essays and comparing them to the MICUSP papers. The intermediate and advanced-level teachers were interested in seeing if there was any notable difference between their students' writing and the papers in MICUSP. I haven't checked with any of them to see if they have had a chance to do this yet, but I think that there will most definitely be a difference between the papers because MICUSP handles academic papers at the collegiate level, while the papers at my institution are mostly written by students who hope to get into the GED program or apply for citizenship. But, it would still be interesting to see what the MICUSP papers have that the ones written at my school don't have.

My first thought on using *Text X-Ray* in the classroom was as a sort of self-check device that the students could use in our technology room. My class is for beginners and we do go over the basic parts of speech, so by having students enter a pre-selected text into the program, having them pick out the nouns in the passage, and then checking their own accuracy with *Text X-Ray* is an excellent way to get the students more engaged with their language learning. Another feature of *Text X-Ray* that I could see myself using in the future for vocabulary purposes is the word list tab. Approximating word meaning from context is a very difficult task in any writing classroom, but if I were able to create a list of words that I think will be difficult for the students in my classroom from a specific passage of written text, and then use *Text X-Ray* to highlight the words in the passage, it would bolster class discussion of the context in which the words are used.

JX, Visiting Scholar—China

Using *Text X-Ray* can highlight how native speakers of English use certain language forms, vocabulary items, and expressions. It offers students the use of authentic and real-life examples when learning writing which are better than examples that are made up by the teacher. It allow students to learn useful phrases and typical collocations they might use themselves as well as language features in context which means that students learn language in context and not in isolation. And it can help students get a broader view of language by comparison. By doing so, students become aware of lexical chunks that are useful when it comes to completing writing tasks. It helps teachers to demonstrate how vocabulary, grammar, idiomatic expressions and pragmatic constraints with real-life language.

JH, ESL Teacher—Korea

What can it do to help students and teachers in the writing classroom? Compared with concordancers, it is VERY user-friendly. I thought that

I could use concordancers only when I prepare the class, but I thought I won't recommend students to use this kind of program before I saw *Text X-Ray*. The program is colorful, and it is very easy to use. With tagging applications, students can easily find the nouns, verbs, and adjectives in their essays. I can use it when I teach verb valency patterns. Since my interest is in teaching grammar using corpus tools, I have been thinking about applicable methodology that I can follow for classroom research on grammar teaching with a tool like this. Because of this visual recognition or representation of "grammar" on the screen, I think students' learning will last more than just from simple rote memory.

MM, Instructor of Japanese, GSU Department of Modern and Classical Languages

Articles are hard to learn for Japanese learners of English since Japanese does not have articles. Texts with highlighted articles (*a*, *an*, *the*) can be used in the writing classroom as a focus on form activity. Compare with MICUSP shows the comparison of frequency of major POS between the corpus and the current essay. By focusing on article use, for example, the program gives a clue to Japanese students of English if they supplied the necessary articles or not. If their frequency of articles is much lower compared to a model corpus, they can focus on their use of articles when they proofread their essay. Word Cloud – it might help students with writing a summary of a text. I remember when I was an undergraduate student, writing a summary in English was so difficult for me. Visual presentation through word clouds can be useful.

CM, IEP Instructor, GSU

I can imagine *Text X-Ray* being very helpful to advanced EAP students who are practicing genre analysis, especially as more and more ELT writing instructors are attempting to empower students to become their own investigators of genre. The *Text X-Ray* tool would allow such students to determine for themselves the differences in, say, nominalization, between academic texts and other types of writing.

In my experience, because of the tendency to associate writing skills with reading skills, a good deal of literacy practice in EAP programs is focused primarily on writing. Even though students may be reading a good deal for homework, there is little explicit instruction on how to approach a text or improve one's reading fluency and/or accuracy. Having taught an upper-level reading course in an ESL program in the past, I certainly would have devoted class time to having students explore their assigned reading through *Text X-Ray*. For example, I may have begun by having students highlight the nouns and do a quick scan for nouns they already know (good

for developing their scanning / skimming skills, as well). Which nouns do they recognize? Which are unfamiliar? Which come from verbs? 4人目

A3.2.3 Other Visualizers and Tools

As noted previously, popular online tools such as *LexTutor* and *WordandPhrase.Info* have components that visualize vocabulary features from texts, providing support for students in discovering patterns of actual language use. In Section C4.2, Roberts uses *WordandPhrase.Info* to show lexical items as categories of frequency with discipline-specific word lists. She presents an IEP lesson at an aeronautical university with authentic content from a single subject area to teach academic English. Nelson in Section C4.4 discusses *LexTutor* and its suite of corpus- and frequency-based tools for English and French language learners and teachers, especially focusing on lexical development, EAP, and CALL. *LexTutor*'s visualized features can show relationships between words (through word lists, word families, concordances, collocations, etc.) or frequency of words and word families. The *VocabProfile* tool is designed to analyze vocabulary use, and other applications include flashcards for learning vocabulary and a hypertext-builder for readings linked with concordances and a *WordReference* dictionary. The following subsections identify additional visualizers and useful programs such as *Sketch Engine*; tools used to visualize online language, especially tweets; and visualizing the language of hip-hop.

A3.2.3.1 Sketch Engine

One of the currently buzzed-about online corpus tools is ***Sketch Engine*** (https://www.sketchengine.co.uk/), which is corpus manager and analysis software developed by the late Adam Kilgarriff and Pavel Rychlý. *Sketch Engine* has evolved in the past few years from its initial online presence in 2013 as a word sketch generator to its present, very impressive applications consisting of three main components: a large database management system, a web interface search program, and a very useful web interface for corpus building and management. These are products that users can access for academic and non-academic license fees. A free 30-day trial is available. It is certainly more than just a visualizer, with its present multiplatform structure, and I do recommend that teachers explore its various features, especially how it can be used for corpus collection and management.

Sketch Engine's database management system (called Manatee) was developed for effective indexing of large corpora. Corpus Query Language (CQL) allows users to easily extract word and phrase-level data from corpora. The current list of available corpora in the program is impressive, and it continues to grow in number, types, and languages. An earlier iteration of the program's word sketch feature (Figure A3.12) provides word distributions obtained from a source corpus like the BNC. In Figure A3.12, *work* is visualized as a noun (alternative POS, *verb* is also an option),

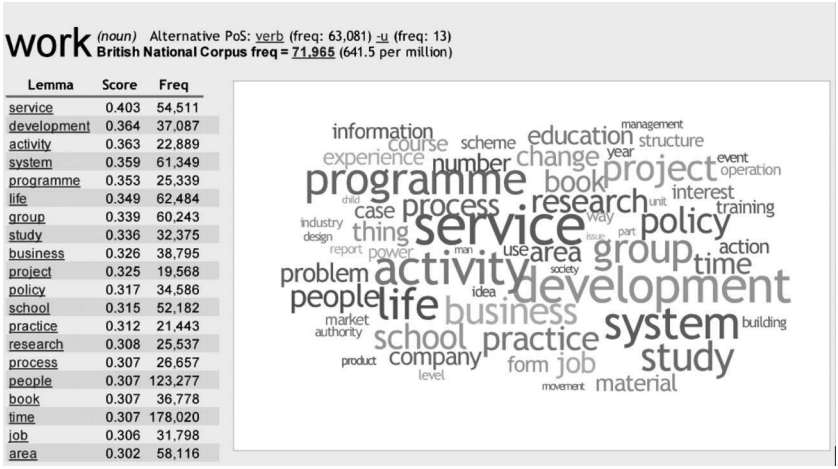

FIGURE A3.12 Sample *Sketch Engine*'s word sketch feature for *work* from the BNC.

appearing 641.5 per million words. The top lemmas are listed and a word sketch, similar to a word cloud, is provided in various colors and font sizes.

A new addition to *Sketch Engine*, updated from Figure A3.12, is the *Sketch Engine for Language Learning* or SkELL tool, intended for teachers and students of English. SkELL allows users to check how a word or a phrase is actually used in a corpus by native English speakers (e.g., from the BNC). All text extracts, collocations, or synonyms are identified and provided automatically by the program. The SkELL tool is free and there is no registration required.

A3.2.3.2 Visualizing Online Language

In the field of *sociolinguistics*, **dialectology** research has benefited from corpus-based quantitative data that support the analysis of dialect variation. For example, Grieve's (2016) study of regional variation in American English from a corpus of newspaper letters to the editor, collected from over 70 cities from across the US, shows how corpus-based methodologies and visualization techniques can be applied directly to researching and teaching regional variation in written language. Grieve carefully designed his corpus to account for geographic regions in the US and their potential influence on variation across a range of linguistic features. One of his primary outputs is a new dialect map of the US showing 12 different dialect regions identified by clusters of linguistic features distinguishing one region from another.

Visualizing online language, especially from social media discourse such as Facebook and Twitter status updates and tweets, has been featured increasingly in several publications and academic articles. Popular culture references, from the broad topic of the language of social media to more specific ones such as President Trump's Twitter analytics, have been explored quite extensively.

Trump, with close to 42 million followers in early 2018, has tweeted an average of 5.4 times per day since he became the US president. His top 10-word list from 2016 includes *Hillary*, *#trump2016*, *crooked*, *Clinton*, *#makeamericagreatagain*, *people*, *America*, *Cruz*, *bad*, and *Trump*. Twitter data are very useful, not only for linguistic analysis but also especially for business and big data analytics. Product sentiment analysis, movie box-office projections, and trending issues or topics are all relatively easy to extract in real time from Twitter using its application programming interface (API). Unlike Facebook, which can be set by users to be exclusively private, Twitter defaults as a public platform.

Eisenstein et al. (2010) used computational models to identify regional markers from user postings on Twitter. For corpus-based dialectology research, the important link here is how internet and mobile technology can code for variables such as **location** in tweets. As it is, Twitter can access users' geographical coordinates from, for example, mobile devices that are enabled by Global Positioning Systems or GPS. This feature produces 'geotagged' text data that researchers can obtain from online logs. Most users' tweets are geotagged, which means that analysts are able to identify the users' locations, especially if they tweeted from their mobile phones. Posts from desktop computers or permanent computer terminals may be identified from their internet access addresses or universal resource locators (URL). There are more and more studies that mine geotagged data online, focusing primarily on trends and internet user traffic. These types of information are useful to marketing analysts and survey companies that collect quantitative tracking data of user behavior from the internet.

One of Grieve and his collaborators' ongoing projects is to document **lexical spread** on Twitter. They are in the process of compiling a multi-billion-word regional monitor corpus, using the Twitter API, consisting of nearly every geocoded Tweet from the US and the UK since 2013, totaling approximately 25 million words per day. Given this large number of geocoded and time-stamped Tweets, it is possible for Grieve and his team (2014) to identify newly emerging words and map their geographical spread over time. An earlier study that they conducted during the first three quarters of 2013 explored "rising" or increasingly prevalent words identified from a particular period: for example, from day 1 of 2013 to day 250 (from January to September). They first extracted 60,000 words that occurred at least 1,000 times in the corpus and identified rising words by correlating word relative frequency per day to day of the year using a Spearman's rank correlation coefficient. Their list of rising Twitter words includes *rn* (right now), *selfie/s* (a photo of oneself), *tbh* (to be honest), *literally*, *bc* (because), *ily* (I love you), *bae* (babe, baby), *schleep* (sleep), *sweg* (swag, i.e., style), and *yasss* (yes). On the declining side, the following are 2013's "falling" Twitter words: *wat* (what), *nf* (now following), *swerve*, *shrugs* (*shrugs*), *dnt* (don't), *wen* (when), *rite* (right), *yu* (you), *wats* (what's), and *yeahh* (yeah).

They were also able to visualize the data per word and then trace the spread of each word across the US For example, for the world *selfie* (named "Word of the Year" for 2013), the following graph (Figure A3.13) clearly shows its

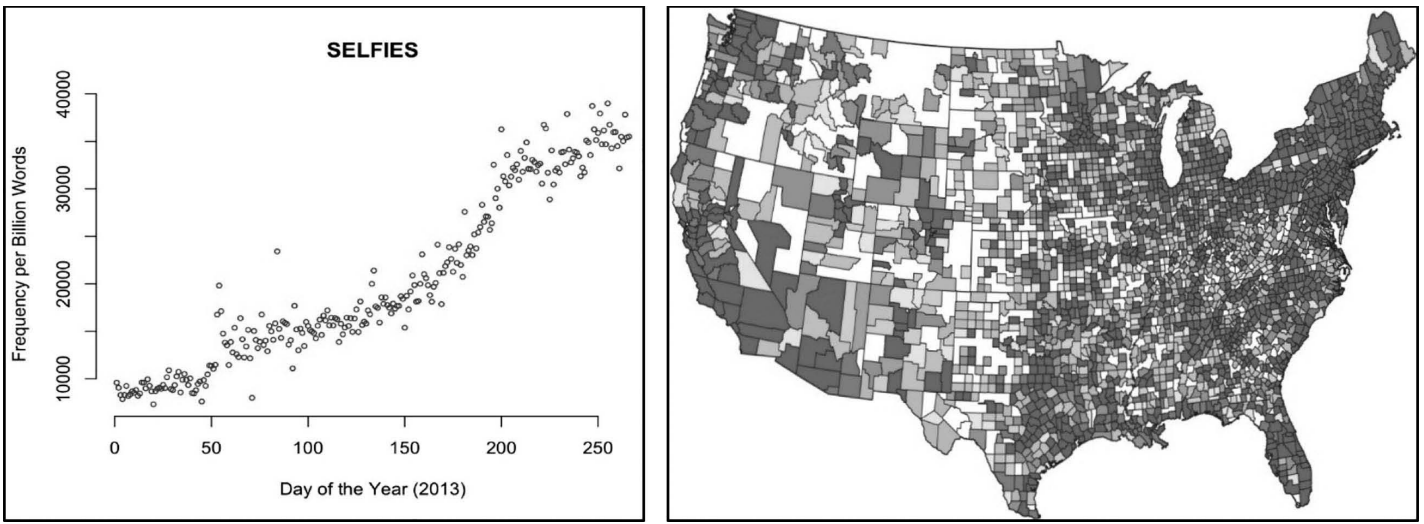

FIGURE A3.13 *Selfie/s* first appearance and dramatic increase in usage from Twitter in 2013. The gray parts in the US map (typically major cities in the Northeast and the Southwest) indicate that *selfie/s* originated from and was immediately popularly used in many major US cities (Grieve et al., 2014).

dramatic linear increase in usage from day 1 to day 250. A 'heat map' from the geotagged tweets shows parts of the US where *selfie* was included as part of a tweet. Additionally, they also visualized the first users of these words in tweets, obtaining profile pictures of Twitter users that could be qualitatively explored according to gender, age, and other variables.

FOR THE TEACHER

The previous figures can be used to initiate small group discussions in a *Sociolinguistics* or *Language in Society* class. Grieve et al. (2014) found that most rising words on Twitter follow an s-curve when presented graphically. They also found patterns: (1) Acronyms were on the rise, but creative spellings were on the decline, (2) there were relatively clear southern and northern US patterns of lexical spread on Twitter, and (3) lexical innovators appeared to include young black women in the South and young white men in the West and the North. (This observation was derived from an examination of profile pictures of Twitter users.) Students examining visualized geotagged Twitter data might be asked to consider and discuss questions such as the following:

- What are your immediate impressions about the data? What jumped out? What are lessons or takeaways from this visualized data?
- Explain what the data/figures and excerpts are about. Answer the question, "So what?"
- Remind students that CL is a research approach, a way of thinking about language that shines the spotlight on language use. What then is a **word?** (Note that Grieve and colleagues referred to *rn*, *ily*, and *yasss* as "words" from Twitter.) What are the implications of these new Twitter words in the study of languages?
- If CL allows investigation of language choice, could we explain why Twitter users prefer a particular word or grammatical form rather than alternatives?

5人目

TAGSExplorer (https://hawksey.info/tagsexplorer/) is a Twitter archive visualization tool developed by Martin Hawksey, which makes use of Twitter to collect tweets related to a particular event or issue hashtag to enable participants to share and contribute relevant comments or responses. As more and more participants tweet, the utilization of the event/issue hashtag will continue to increase and can, thereby, enable greater public visualization of archived tweets. Users can interact with their own and other users' tags, and the visualization could be shared online. Figure A3.14 shows an example of this "queryable visualization" format, which was made possible by using Google Sheets as a database of tweets.

残り3人目がもう1回

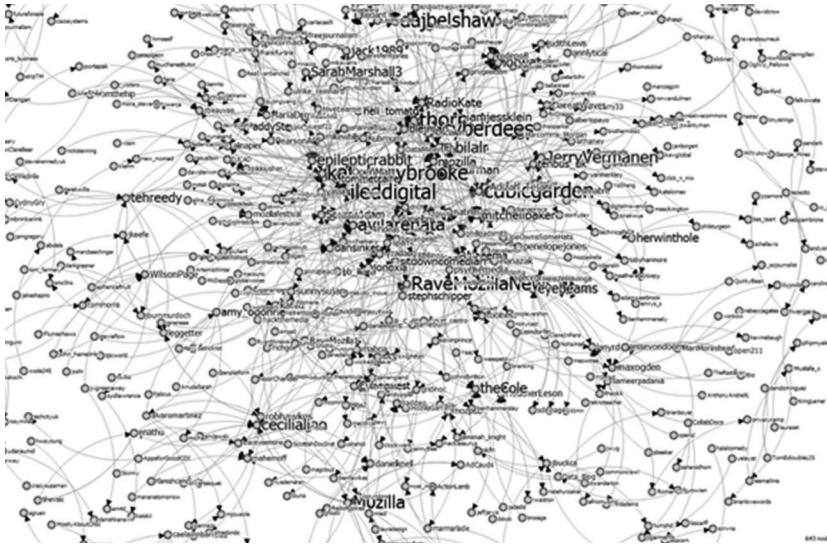

FIGURE A3.14 Representation of *TAGSExplorer*'s nodes of event hashtags and interactive visualization.

What does it all mean? Hawksey's goal in developing *TAGSExplorer* was to archive event hashtags and create an interactive visualization of the conversation threads on Twitter. The device makes use of many data points identified by users and provided in real-time by Twitter. Twitter, in this context is, therefore, an automatic corpus available for instantaneous analysis. Although it may not be immediately applicable to classroom teaching or in the teaching of linguistic features aside from the tracking of vocabulary use in this register, the tool is a reflection of language development online and how emerging technologies are providing linguists and teachers access to authentic texts. *TAGSExplorer*, in the future, could be the model for individualized, learner-centered instruction on language description. This may first start with vocabulary instruction, followed by focus on form/structure activities for learners to enable them to be more aware of sentence-level nodes of language as they interact with the tool. Learners will be asked to focus on discovering unique patterns of language that they can use in their own writing, whether sending tweets or, potentially at some point in the future, writing more formal academic essays.

A3.2.3.3 Exploring the Language of Hip Hop (also Hip-Hop)

Finally, although they do not typically refer to CL as their underlying method of analysis, recent projects developed by researchers associated with the Rap Research Lab and similar groups such as The Hip-hop Archive & Research Institute at Harvard University, and The Frank-Ratchye STUDIO for Creative

Inquiry at Carnegie Mellon University utilize a corpus of hip hop lyrics to explore vocabulary use in hip hop and to visualize and compare artists' creative use of language. Hip hop has been a leading source of linguistic innovation and has also now been studied across academic fields in the digital humanities, media criticism, and data visualization. In many of these academic studies, the language of hip hop is viewed as a cultural indicator. Tahir Hemphill has developed a searchable rap almanac the *Hip Hop Word Count* (http://staplecrops.com/), which examines hip hop lyrics and allows a visualization tool to draw out shapes and circular lines from the lyrics, revealing a layer of creative work and the aesthetic focus that artists pursue in their songs.

The *Hip-Hop Word Count* is a searchable ethnographic database built from the lyrics of over 40,000 hip hop songs (and growing) from 1979 to the present (Hopkins, 2011). From this database, linguistic details of hip hop songs can be explored and compared. As Hemphill suggests, these data can then be used to not only derive interesting statistics about the songs themselves, but also potentially to describe and explain the culture behind the music. An illustrated visual on the artist, a particular song, and linguistic information such as total words, average syllable per word, average letters per syllable, average letters per word, polysyllabic words, and finally education level (e.g., some high school or high school graduate) and readability or reading level are provided. Reading levels are identified as "Readers' Digest," "Time Magazine," and others. In the following comparison data, adapted from Hemphill's site, Kanye West's "Big Brother" and Tupac Shakur's "Trapped" received word count scores of 9 and 12, respectively. (The higher the word count, the more sophisticated the lyrics are, arguably.) Linguistic metrics of the two songs are provided (Figure A3.15).

How can analyzing hip hop lyrics teach us about cultures or subcultures? Song-level comparisons are potentially interesting to students, especially those who like this genre of music, but they can also apply this approach to other genres or extend the comparison to two or more corpora. For example, my students have always been curious about the differences of vocabulary use and themes between country music and hip hop. They explore the distribution and functions of words like *love*, *God*, *freedom*, *America*, and tagged POS features such as personal pronouns, verb tenses, passive structures, and nominal modification. The *Hip Hop Word Count* also provides time and geographic location identifiers based on where the artists came from in the US and related comparisons of metaphor use and other figures of speech, cultural references, phrase and rhyme style, meme and socio-political ideas. Hemphill's database then converts various data points into an explorable visualization frame that charts "migration of ideas and builds a geography of language."

Daniels (2017) used a token analysis method—basically, a type-token ratio—to determine hip hop artists' vocabulary range, identifying unique words from an artist's first 35,000 song lyrics collected in the corpus. His various results allowed him to create a master list of who has the most to the least unique and

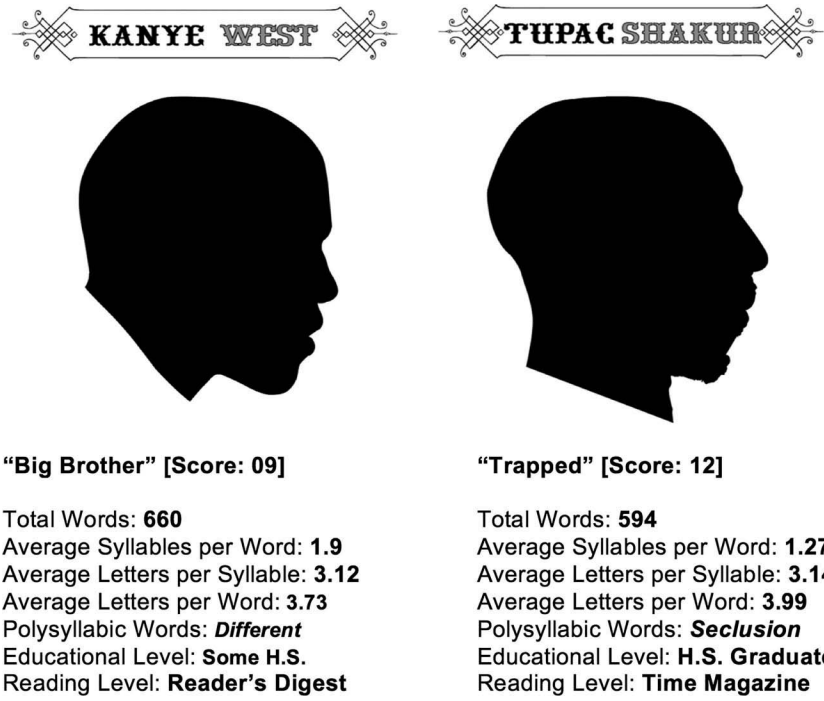

FIGURE A3.15 Comparison of "Big Brother" (Kanye West) and "Trapped" (Tupac Shakur) from the *Hip Hop Word Count*. Adapted from Hemphill: http://staplecrops.com/.

diverse vocabulary range. An online interactive visualizer (https://pudding.cool/2017/02/vocabulary/index.html) provides a set of data that also show how artists compare to Shakespeare and Herman Melville's *Moby Dick*. Results from this approach revealed that Aesop Rock (born 1976), based in Portland, Oregon, was the artist with the "largest vocabulary in hip hop," with 7,392 unique words used. By comparison, Shakespeare's total was 5,170; Melville's was 6,022, based on the first 35,000 words of *Moby Dick*. The artist with the smallest vocabulary was DMX (born, 1970, from New York), with 3,214 unique words. Daniel's comparison chart shows that a majority of artists plot in the 3,600–5,000 range, below Shakespeare. There were not many women in the dataset; Lil' Kim (born, 1974, from New York), with 4,470, and Nicki Minaj (born, 1982, from Trinidad and Tobago, raised in Queens, NY), with 4,162.

Note

1 We plan to launch a full, new version of *Text X-Ray* upon completion of our usability tests. If you want to access the beta version, please send an email to **textxray.beta@gmail.com** for instructions on how to run the program online.