Linguistic Research 36(1), 1-34
DOI: 10.17250/khisli.36.1.201903.001

The advantages and challenges of "big data": Insights from the 14 billion word iWeb corpus*

Mark Davies · Jong-Bok Kim*****
(Brigham Young University · Kyung Hee University)

Davies, Mark, and Jong-Bok Kim. 2019. The advantages and challenges of "big data": Insights from the 14 billion word iWeb corpus. *Linguistic Research* 36(1), 1-34. The iWeb corpus contains nearly 14 billion words from 22 million web pages, and it has been designed in a way that allows users to quickly and easily create "Virtual Corpora", in order to focus on websites that are related to their areas of interest. The data from this very large corpus provides very detailed information on syntactic, morphological, lexical, and semantic phenomena, in ways that would never be possible with a small 100 million or 500 million word corpus. In addition, the corpus provides a number of features that are not available with other large corpora, such as the ability to perform advanced searches of the top 60,000 words in the corpus, and to see a wealth of information on each of these words – definitions, links to images and audio, translations, detailed frequency information, related topics, collocates, word clusters, re-sortable concordance lines, and much more. Finally, we discuss the challenges of large corpora, and how the corpus architecture that is used for iWeb has uniquely been designed to address these challenges. **(Brigham Young University · Kyung Hee University)**

Keywords iWeb, virtual corpora, big data, BNC (British National Corpus), COCA (Corpus of Contemporary American English)

1. Introduction

Advances in technology have made possible very large corpora that would have been unthinkable even 10-15 years ago. With access to the right hardware and software, it is now possible to scrape billions of words of data from the Web and create a corpus that can be used to research a wide range of linguistic phenomena. As we will see, this extremely rich data can then be used to answer

* We thank anonymous reviewers of the journal for the constructive comments and suggestions. This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2014S1A2A2028437).
** First author
*** Corresponding author

basic questions about language variation, in ways that would be quite impossible with smaller 100 million or 500 million word corpora.

This paper will deal with the iWeb ("Intelligent Web") corpus, which was created as part of the BYU suite of corpora, and which was released in mid-2018. Section 2 briefly discusses the composition of the corpus and the steps that were followed to create the corpus. Sections 3-4 provide several examples of how the rich data from iWeb can be used to examine syntactic, morphological, lexical, and semantic phenomena that cannot really be studied with much smaller 100 million or 500 million word corpora. Section 5 discusses why size is not everything, and why it is also necessary to take into account the corpus architecture, to create large corpora that are actually usable. Section 6 shows how data from large corpora can provide useful "word-level" information, which can be used for teaching and learning. Finally, Section 7 summarizes the main findings of the paper.

2. Creating the iWeb corpus

There are three sets of very large, 10+ billion word corpora. The first are the Sketch Engine corpora, which are available for many different languages (https://www.sketchengine.eu). The second are the "Corpora from the Web" corpora (https://www.webcorpora.org). The third corpus that is 10 billion words in size or larger is the iWeb corpus, which was released in mid-2018, and which joins several other billion word corpora from corpus.byu.edu.

The iWeb corpus contains about 14 billion words in 22,388,141 web pages from 94,391 websites. Unlike other large web-based corpora, iWeb was created by focusing on particular websites, rather than just scraping data from web pages on random websites. The following are the steps that the first author followed to create the iWeb corpus which was created by the first author alone.

1. Data were downloaded from Alexa.com (created by Amazon) on the top 1,000,000 websites from throughout the world, and their list is based on the number of users of these websites.

2. For each of these one million websites, the Alexa data were used to find what percentage of the users are from the US, Canada, Ireland, the UK, Australia, and New Zealand. The idea was to use websites that would mainly be in English, as opposed to websites from India or Nigeria or Singapore (or obviously China or Japan or Russia), where they might contain material from other languages.

3. For each of the top 200,000 websites (from step #2), the URLs from searches on either Google or Bing were obtained (and stored in a relational database). The author basically just searched for web pages containing the word "of" from each of these websites, and (because nearly every page will have the word "of"), Google and Bing yielded "random" web pages from each of these websites. Because Google will block repeated queries from the same IP address (somewhat less of a problem with Bing), these searches were very slow and methodical (to "stay below their radar"), and it took approximately three months to get all of the 27,000,000+ URLs.

4. All of the web pages for each of these websites were then downloaded, using custom software written by Spencer Davies in the Go programming language. It took about three days (using five different machines) to download the (approximately) 27 million web pages (at about 100 pages per second).

5. Approximately 30,000 of the 200,000 websites (from #3) were eliminated from the corpus, because of one of the following:

▪ These were what might be called "transaction" websites, where there is little if any publicly-available data from "static" web pages. Examples might be VPN sites, torrent sites, or sites that require users to log in or to do a specific search to see pretty much anything else. For example, think of Google itself. 99.9% of anything valuable from Google will be the results for a specific search, not a static web page at www.google.com. So a list of random URLs (for static web pages) from www.google.com would not be very useful.

- Websites that were blocked by the proxy server at BYU. The vast majority of these were porn sites, although there were a few for gambling, "hate speech", proxy avoidance, and other "blocked" sites.

6. JusText was used to remove "boilerplate" material (headers, footers, sidebars, etc), and then each of the pages was tagged with the CLAWS 7 tagger.

7. At this point, there were about 170,000 websites. To obtain websites that had enough words and web pages to get a good sampling of the language from the website, the first author set a minimum threshold of 10,000 words in at least 30 different web pages from each website, and this eliminated another 65,000 websites, leaving about 105,000 websites.

8. Repeated tests and procedures were then performed to find duplicate web pages and phrases. The author searched for duplicate n-grams (primarily 11-grams), looking for long strings of words that are repeated, such as "*This newspaper is copyrighted by Company_X. You are not permitted...*" (= 11 words, including punctuation). The author ran these searches many times, in many different ways, trying to find and eliminate duplicate texts, as well as duplicate strings within different texts. Because there is a great deal of duplicate material on web pages (even after running programs like JusText), this eliminated approximately 10,600 more websites that had less than 10,000 words or 30 web pages.

9. After all of these steps, the final output was 94,391 websites (each with a minimum of 10,000 words and 30 web pages), for a total of about 14 billion words from 22,388,141 web pages.

In other large web-based corpora, the websites are essentially "random", and the vast majority of websites might contain just a handful of web pages that were scooped up as the list of URLs was generated from links on other pages, and which means that users cannot search by the website itself. But because of the systematic, principled way in which the websites were selected for iWeb, there is an average of 245 web pages and 140,000 words for each of the 94,391 websites.

And because of the underling architecture of the corpus, users can quickly and easily create Virtual Corpora to search by website, and to find the websites that refer the most to a particular word, phrase, or even topic, which is discussed more in Section 5.

3. The advantages of very large corpora for syntax and morphology

Very large corpora – such as iWeb – can provide insight into linguistic variation that would not be possible with smaller 100 million or 500 million word corpora. In this section we will consider a number of syntactic constructions where the large amount of data from iWeb allows us to look at the interaction of syntax, semantics, and pragmatics in ways that probably would not be possible with these smaller corpora.

First, consider “auxiliary stacking”. English can have three auxiliaries (four, if modals are added), as in examples like the following from iWeb:

(1) a. some managers **have been being trained** to act as guards.
b. Let’s face it, problems **have been being solved** as long as man has been faced with them.
c. This tried and true line of car care products **have been being used** by the professionals for years.
d. but I **may just have been being** a bit of a bratty kid.
e. The result **would have been being sent** to do forced labor in Siberia.
f. It **could have been being** physically, sexually, emotionally or verbally **abused** by an adult.

This is quite an infrequent construction in English, in large part because the main verb has to be amenable to use with the passive and the perfect and the progressive, and there needs to be some situation in which all three of these modality, aspects, and voices are all important at the same time.

There are only two tokens with three consecutive auxiliary verbs (leaving aside modals) in the 100 million word British National Corpus (BNC), both for different types (distinct strings of words). In the Corpus of Contemporary

American English (COCA), there are 16 tokens with 15 different types. But because of the relatively small number of tokens, none of these types have three or more tokens, which means that it is hard to determine which types of main verbs would occur most with this construction. For example, is there something in terms of the event structure of the main verb that the following have in common: *have been being used, have been being trained, have been being thrown, have been being skinned, have been being set, have been being handled,* and *have been being charged*? With so few tokens for each type (2 for the first string and 1 for all of the others), it is quite difficulty to detect any patterns.

But in iWeb there are 637 tokens, 384 types, and 51 of these 384 types have three tokens or more. The most common strings are *has been being used* (25 tokens), *have been being used* (24), *have been being made* (15), *have been being treated* (13), *has been being worked* (8), *'ve been being treated* (7), *have been being paid* (7), *have been being built* (6), and the following strings that occur five times each: *'ve been being told, has been being built, has been being groomed, has been being played, have been being attacked, have been being printed,* or *have been being taken*. Once we have a list of these strings that occur multiple times, we can begin to ask questions about the underlying semantic composition of the strings. For example, is there something special about the events expressed by *use, treat*, or *build*, which would cause them to be so much more common in a construction that involves perfect + progressive + passive? We will leave it to others to consider this question, but the point is that without a large amount of data from a very large corpus, this is the type of issue that we could not otherwise even begin to consider.

A second example is the "into VERB-ing" construction, shown in the following examples from iWeb:

(2) a. kidney disease can **fool you into thinking** things are ok for a while.
b. but don't let it **force you into buying** a more expensive flight.
c. he **lured me into making** an application but I never got even to the interview stage

There are at least 15 different verbs that occur in iWeb that do not occur even once in the BNC (100 million words), COCA (560 million words), or Corpus of Global Web-based English (GloWbE; 2 billion words), including *stress* (18 tokens

in iWeb), *enroll* 16, *break* 11, *rationalize* 10, *wow* 10, *punish* 9, *pride* 9, *raise* 8, *instigate* 8, *bias* 8, *engineer* 8, *antagonize* 7, *manifest* 7, *exalt* 7, *solicit* 7, *plug* 6.

But what does it matter that the construction occurs with certain verbs in iWeb but not in the other corpora? The importance is that some of these new verbs can signal more general semantic shifts that are taking place. For example, Kim and Davies (2016) and Davies and Kim (2018) show that in the 1800s and the first half of the 1900s, virtually all of the verbs were "negative" verbs like *fool, force*, or *lure* (shown above), or at the very least "neutral" verbs like *lead* (*and I kind of* ***led her into saying*** *no*). Just within the last 30-40 years, however, extremely rare cases of "positive" verbs have begun to appear, which suggests an interesting semantic evolution of the construction.

Due to its large size, we find a number of similar cases in iWeb. A number of different positive verbs occur at least six times in the corpus, as in the following examples with *wow, pride*, and *enthuse*. There are probably another 30-40 distinct positive verbs that occur with "into VERB-ing" in iWeb at least once.

(3) a. fans of precious series will usually give reboot a chance to **wow them into liking** it.
b. If Samsung is trying to **wow us into trusting** it again, it's done a pretty good job
c. We **pride ourselves into putting** good people into a great business.
d. These three things are something that I **pride myself into proving** and achieving by going above and beyond to exceed your expectations.
e. Phil took on the role … which has **enthused him into moving** forward with the latest models.
f. Peter had **enthused them into thinking** they had seen the Master.

Another very recent shift with the construction is what we have called the "indirect causative". In a typical case like *John talked Mary into going to the movies*, there is a fairly "tight" semantic link between John's talking Mary about something and her doing it. But just within the last 10-20 years, very sporadic cases have begun to emerge in which the pragmatic linkage is much less direct. Consider the following from iWeb:

(4) a. They also automatically **enrolled me into receiving** and charging me for another product.
b. my #1 strategy for talking to potential clients so they practically **enroll themselves into working** for you
c. Hopefully this policy **manifests itself into retrofitting** protected bike lanes.
d. behavioural issues which a lot of times **manifests itself into using** drink or drugs
e. given the unwillingness of the much-vaunted Rivaldo to **raise himself into threatening** us in the slightest
f. when the lower part of him has become so evolved that it **raises itself into becoming** at one with the higher part of himself
g. that now has **adapted employees into becoming** an integral part of marketing
h. the mind … has **adapted itself into remembering** important events

In iWeb we are able to very quickly find four different verbs that allow for this more indirect reading, each with six tokens or more. And there are probably another 20-30 different verbs that have just two or three tokens, and perhaps hundreds of other "indirect causation" verbs that occur just once. Compare this to COCA, where there are only 5-6 tokens altogether, and none in the BNC. Again, the rich data from iWeb provides supporting evidence for very recent, very low-frequency changes that may signal more general shifts with the construction.

A third construction where there is interesting data from the large iWeb corpus is with the "way" construction (see Israel 1996 and Goldberg 1997).

(5) a. **Make your way through** the castles by picking up items.
b. I'm **working my way through** those massive piles of book.
c. all the potential elements that would **find their way into** our story
d. One of the best and least known features which has **made its way onto** the Galaxy S8

It would presumably be helpful to have more than just a few tokens with a

given "*way* construction" verb, in order to discover how idiomatically the construction is being used. Assuming at least 10-20 tokens with a given string (e.g. *make your way through* or *find their way into*), we find that there are 47 strings with a frequency of 10 tokens in the BNC, and only 11 strings with a token frequency of 20. In COCA it is 328 strings (10 tokens) and 140 strings (20 tokens), respectively. But it is of course much richer in iWeb, where there are 3,683 distinct strings that occur at least 10 times, and 1,999 strings for 20 tokens.

In terms of the ever-extending boundaries of the "way construction", the most interesting strings are those that occur in iWeb, but not in smaller corpora like the BNC and COCA. For example, there are 361 different strings that occur at least 10 times in iWeb, which do not occur even once in the 560 million word COCA corpus. These verbs include the following: *network* (146 tokens in iWeb), *cruise* 144, *print* 116, *roar* 104, *travel* 93, *solve* 87, *putt* 74, *mash* 71, *inflate* 69, *browse* 67, *dive* 66, *traverse* 66, *rationalize* 65, and *blog* 65, e.g.:

(6) a. start **networking your way into** the industry or field that you want to move into
b. Greece could leave the Euro and attempt to **print its way out** of excessive debt.
c. The best part about the game is **solving your way through** the tense scenarios.
d. if you decide to **dive your way into** finding hair growth products
e. Secrets for **Blogging Your Way to** a Six-Figure Income

A fourth syntactic construction involves the contrast between [to V] and [V-ing] complements (see Rohdenburg 2009, Vosberg 2003, Mair 2002, and Rudanko 2000). In addition to looking at the overall "macro-level" shift towards [V-ing] complements over time, we can also focus on "micro-level" shifts with particular verbs (e.g. *start to walk / start walking*; *love to watch them / love watching them*) and adjectives (e.g. *crucial to understand the problem / crucial to understanding the problem*). For example, Rudanko (2012) focuses on the one particular adjective *prone* to examine minutely how the shift from [to V] to [V-ing] has been spreading over the last 20-30 years.

In a "small" 100 million word corpus like the BNC there are 154 tokens of

[to V] and 70 tokens of [V-ing] with *prone*, and there are 862 tokens of [to V] with *prone* in COCA and 552 with [V-ing]. This may seem like enough tokens, but the problem comes when we start looking at the individual verbs with which *prone* occurs. If things are slowly shifting from [to V] to [V-ing], then it may be that this is slowly spreading from one type of verb to another (for example, a particular event structure. But in this case, we might be looking at 100 different verbs, and now the 224 token tokens from the BNC are far too few to compare what is happening with the different verbs.

In the case of iWeb, however, we have extremely rich data. There are 22,965 tokens of [to V] and 22, 949 tokens of [V-ing] with the one single adjective *prone*. This large number of tokens means that we *can* in fact compare what is happening with a particular verb, such as *prone to develop / developing*, *prone to break / breaking*, or *prone to fall / falling*. This large number of tokens is also useful when we start examining the "verbs" one by one, and discover that a great many of them are actually nouns that have been mistagged as verb (e.g. it is *prone to rust / wear / dry* (*rot*) / *rot*).

Table 1 shows the frequency of [to V] (*prone to develop*) and [V-ing] (*prone to developing)* with the top eleven verbs in iWeb, and then the corresponding numbers for COCA and the BNC.

Table 1. prone + [to V] and [V-ing]

prone +	iWeb to_V	iWeb V-ing	% iWeb V-ing	COCA to_V	COCA V-ing	% COCA V-ing	BNC to V	BNC V-ing
get	525	**1505**	74%	10	**19**	66%	1	
develop	754	**1292**	63%	18	**20**	53%	4	
make	374	**704**	65%	17	**23**	58%	4	3
break	335	**889**	73%	6	**9**	60%		1
fall	196	**411**	**68%**	1	**12**	92%	3	1
cause	175	**178**	**50%**	6	2	25%		
believe	136	67	33%	8	1	11%	1	
give	117	**171**	59%	6	6	50%	1	1
go	117	**182**	61%	8	3	27%	1	
think	116	65	36%	6	2	25%	2	
forget	113	95	**46%**	3	1	25%	1	1

In a "small" corpus like the BNC, such a study would of course be impossible – there just aren't enough tokens. Even in the 560 million word COCA corpus, there is probably only enough data to look at 4-5 different verbs. The data from Table 1 shows that [V-ing] is more common than [to V] in COCA with *get, develop, make, break*, and *fall*, but some of this is based on just 12-15 tokens with a given verb; 3-4 tokens the other way and we would come up with quite different results. In other words, with the COCA data (and even much more the data from the BNC), we are left wondering whether the data is "noise" that is due to chance. But with iWeb we have hundreds of tokens with each of these verbs, which helps to validate the data from COCA. And in some cases like *cause* (e.g. *prone to cause / causing*) and *forget* (e.g. *prone to forget / forgetting*), the data from iWeb shows that the shift towards [V-ing] is actually twice as advanced as with the fragmentary data in COCA (and virtually non-existent data in the BNC).

As a fifth syntactic construction, consider [BE so not ADJ] as in the following:

(7) a. But do you know what? We **are so Not invisible**!
b. and I **am so not comfortable** with blog reader sites using my content on Pinterest
c. None of this jives and **he is so not spiritual**. How do I know this?
d. I'm neither a vegan nor not drinking. I **was so not pregnant**. But this is unbelievable.

Stange (2017) confirms that intuitions of most native speakers of English, who would say that the construction is very informal, and that it is definitely increasing over time. The BNC is too old and too small to have any tokens of the construction. In COCA there are 77 tokens with 49 different types. This data shows that the construction is definitely more common in the informal genres like Spoken and Fiction, and that it is increasing over time (although the small numbers for 2015-2017 leaves very recent changes a bit tentative).

SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC	1990-1994	1995-1999	2000-2004	2005-2009	2010-2014	2015-2017
30	24	18	5	0	0	4	12	25	28	8
116.7	111.8	117.4	113.0	111.4	104.0	103.4	102.9	102.0	102.9	62.3
0.26	0.21	0.15	0.04	0.00	0.00	0.04	0.12	0.24	0.27	0.13

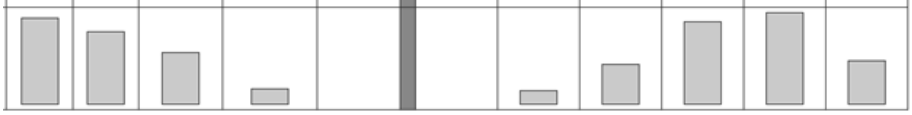


Figure 1. BE so not ADJ construction in COCA

But if the question is *what* adjectives occur most with the construction, COCA may not have enough data. There are only two strings that have a frequency of more than three tokens: *is so not true* and [it]*'s so not true*. There are other adjective like *funny, fair, true*, and *soothing*, but the token counts are very low.

In iWeb, there are 1,379 tokens with 456 types, and most importantly, 98 of these occur three times or more. In total, there are 281 different adjectives that occur with the construction, with the most frequent being *true* (187 tokens), *ready* 87, *alone* 86, *cool* 63, *good* 62, *fair* 56, *only* 43, *fun* 41, *right* 40, *funny* 34, *sure* 32, *interested* 25, *okay* 20, *used* 18, *happy* 17, *surprised* 16, *normal* 15, *scary* 14, *necessary* 13, *sexy* 13, *easy* 12, *perfect* 12, *crafty* 12, and *ok* 12. As can be seen here, nearly all of these adjectives are evaluative or emotive, rather than descriptive or categorial (# *it is so not green*, # *he was so not social*). But only iWeb has the rich data to confirm these intuitions.

The preceding five examples considered syntactic phenomena in which billions of words of data provide insight into the interaction of syntactic and semantic phenomena, in ways that would not be possible with smaller corpora. Let us now consider briefly two examples of how very large corpora like iWeb can provide rich data on morphological variation.

First, consider something as basic as morphological variation with different verb forms. For example, Table 2 shows the frequency in both iWeb and COCA for competing forms of the past participle (e.g. *this has proven / proved*). (Note that the verbs are arranged in ascending order of frequency of the two forms in COCA.) In this table, for example, iWeb has 2114 and 1434 tokens (respectively) for the two forms of *strive*: *HAVE strived* and *HAVE striven*.

Table 2. Competing forms of past participles

Verb	Form 1	Form 2	iWeb		COCA		iWeb #1	COCA #1
saw	**sawed**	**sawn**	**87**	**111**	**22**	**0**	**0.44**	**1.00**
shear	sheared	shorn	327	100	33	17	0.77	0.66
sow	sowed	sown	291	2010	12	93	0.13	0.11
strive	strived	striven	2114	1434	62	77	0.60	0.45
speed	**speeded**	**sped**	**308**	**1510**	**40**	**109**	**0.17**	**0.27**
sew	**sewed**	**sewn**	**198**	**1851**	**29**	**133**	**0.10**	**0.18**
shave	shaved	shaven	1931	71	258	3	0.96	0.99
thrive	thrived	thriven	5032	45	361	0	0.99	1.00
bite	**bit**	**bitten**	**1375**	**3418**	**56**	**339**	**0.29**	**0.14**
swell	swelled	swollen	2205	2422	310	169	0.48	0.65
wake	waked	woken	134	4629	41	491	0.03	0.08
drink	**drank**	**drunk**	**1950**	**4705**	**42**	**550**	**0.29**	**0.07**
shrunk	shrunk	shrunken	8845	385	940	31	0.96	0.97
beat	beat	beaten	4113	19570	351	1371	0.17	0.20
strike	stricken	struck	276	17370	19	1837	0.02	0.01
prove	proved	proven	53306	110034	3531	3280	0.33	0.52
show	showed	shown	8327	293060	492	14956	0.03	0.03

For many of the verbs, the data from COCA is probably sufficient. Most of the verbs (with the exception of *have sawn/sawed* and *have shorn/sheared*) have 100 tokens or more in COCA. But the interesting cases (5 of the 17 verbs) are those where the data from iWeb is quite different from COCA. In the case of the verbs *sew, saw, bit, speed*, and *drink* (bolded in Table 2), the relative percentage of the two forms (the rightmost two columns) is quite different in iWeb and COCA. And in each of these five cases, there are 9 to 13 times as many tokens in iWeb as in COCA, which allows use to be even more confident that the data is accurately modeling what is going on in the language.

Let us consider one other example where iWeb provides rich data on morphology, in this case morphological creativity. Table 3 shows the data for seven different suffixes and prefixes in English. In each case, a handful of interesting forms are shown, as well as the number of types (distinct forms) in iWeb, COCA, and the BNC. In each case, the table shows the total number of types followed by the number of types that occur more than once (which shows that the form is not just a strange "once-off" token in some random text). For example, there are 764 distinct **calypse* words in iWeb, and 297 of these occur

two times or more.

Table 3. Morphological / lexical creativity

	iWeb forms	iWeb (tot / f>1)	COCA	BNC
*calypse	clownpocalypse, memepocalypse, grandmapacalypse, cringepocalypse	764; 297	30; 11	2; 1
*geddon	snowmageddon, rockmageddon, popupgeddon, datamageddon	517; 210	18; 5	4; 2
*fest	sneezefest, jackfest; thirstoberfest, gropefest, donutfest, flopfest	5822; 3160	547; 216	39; 18
*athon	bonkathon, kissathon, relaxathon, subscribathon, slugathon, snoreathon	1426; 647	95; 33	22; 11
*phobia	bathophobia, repairophobia, emotophobia, kinkphobia; pancakeaphobia	2047; 960	189; 70	37; 19
smart*	smartflush, smartlease, smartshade, smartwrap, smartkarma	11203; 5529	556; 237	80; 38
sexual	lumbersexuality, sexualninja, omnisexuals, spiritual-sexual, robotsexual	5269; 1811	619; 207	144; 51

As can be observed here, the difference in the amount of data is quite striking. On average there are 17 times as many types (distinct forms) in iWeb as there are in COCA, and 137 times as many types as in BNC. So while a small 100 million word corpus may provide examples of some creative uses of derivational morphemes in the language, it is in a very large corpus like iWeb that we can truly appreciate the full range of morphological and lexical creativity.

4. The advantages of very large corpora for lexis and meaning

The advantages of very large corpora are not limited to investigations of syntactic and morphological variation. The very large Sketch Engine corpora, for example, stress the advantages in terms of word level phenomena like word frequency and word meaning, and these are the phenomena that we will briefly consider in this section.

Consider first lexis, as seen through the lens of word frequency. Figure 2 is from https://www.wordandphrase.info/new, and these tables show words in COCA (560 million words) at three different frequency levels: near word #12,200,

words near #24,200, and words near #44,200. (Note that the rank order (e.g. #12,200) is actually a function of the raw word frequency, as well as the dispersion, or how well the word is spread across the different sections of the corpus).

RANK #	PoS	WORD	TOTAL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
12200	j	SILVERY	1244	10	738	318	149	29
12201	R	MYSTERIOUSLY	1139	147	372	297	174	149
12202	V	RE-ENTER	1142	129	359	250	243	161
12203	N	BOROUGH	1230	121	159	181	514	255
12204	j	ENIGMATIC	1130	52	245	312	219	302
12205	j	TIBETAN	1219	115	151	239	400	314
12206	j	RESTING	967	0	270	338	141	218
12207	N	ALLURE	1196	97	172	423	344	160
12208	j	DISTRAUGHT	1165	297	417	155	220	76
12209	j	UNRULY	1167	103	393	251	221	199
12210	V	WAG	1220	121	753	177	120	49

RANK #	PoS	WORD	TOTAL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
24282	j	MARSHY	279	11	88	93	54	33
24283	N	HELIX	330	17	51	154	36	72
24284	N	FLAXSEED	363	5	2	332	20	4
24285	N	ALABASTER	290	14	155	68	27	26
24286	j	CARAMELIZED	330	14	21	138	157	0
24287	j	NON-COMPETITIVE	285	13	2	57	48	165
24288	V	ANNOTATE	285	14	52	63	24	132
24289	N	LIBRETTO	289	11	27	40	120	91
24290	j	ENDOCRINE	286	19	9	141	25	92
24291	j	BROKEN-DOWN	271	17	108	45	82	19
24292	V	IMPUGN	279	92	29	42	53	63

RANK #	PoS	WORD	TOTAL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
44242	N	METROPLEX	74	6	5	11	37	15
44243	j	UNSUBSIDIZED	57	2	2	17	16	20
44244	j	AFTER-MARKET	49	7	1	20	9	12
44245	j	HUMAN-INTEREST	57	1	12	8	12	24
44246	j	PROCESSUAL	53	0	0	0	0	53
44247	N	CHILDLESSNESS	55	1	11	9	18	16
44248	j	F-SERIES	50	4	0	17	29	0
44249	j	GAME-LIKE	52	3	1	9	6	33
44250	j	SOLAR-TYPE	49	0	1	31	0	17
44251	M	MID-1998	49	1	0	13	24	11
44252	N	PAROTID	63	0	0	0	0	63

Figure 2. Words near #12200, 24200, and 44200 in the COCA word frequency list

There are a number of interesting points that we could make about the types of words at these three different frequency “strata”, such as the high degree of hyphenated and academic words at the lower frequency level, or the fact that the more frequent words include more verbs and adverbs. But for the purposes of our discussion here, consider the frequency of the words at each of these

frequency levels in COCA. Near word #12,200 they occur about 1200 times each, decreasing to about 300 tokens for words near #24,200, and then about 50-60 tokens for words near #44,200.

Consider now the words near #12,200, #24,200, and #44,200 in iWeb. (These are taken from the "Browse" word listing at the iWeb website).

RANK	FREQ	Word	PoS	Audio	Video	Image
12208	37002	diffusion	NOUN			
12209	36994	parity	NOUN			
12210	36992	three-dimensional	ADJ			
12211	36987	standardize	VERB			
12212	36983	triangular	ADJ			
12213	36981	subside	VERB			
12214	36972	wiper	NOUN			
12215	36971	psyche	NOUN			
12216	36967	acclaim	NOUN			
12217	36951	puddle	NOUN			
12218	36951	skew	VERB			
RANK	FREQ	Word	PoS	Audio	Video	Image
24204	8145	trickster	NOUN			
24205	8144	backroom	NOUN			
24206	8144	overhanging	ADJ			
24207	8142	deadbolt	NOUN			
24208	8141	smallish	ADJ			
24209	8141	thirty-one	NUM			
24210	8139	potash	NOUN			
24211	8139	heli	NOUN			
24212	8138	aerate	VERB			
24213	8138	double-edged	ADJ			
RANK	FREQ	Word	PoS	Audio	Video	Image
44206	1415	icecap	NOUN			
44207	1415	nuthatch	NOUN			
44208	1415	intrapersonal	ADJ			
44209	1415	bronchus	NOUN			
44210	1415	all-tournament	ADJ			
44211	1415	ranchero	NOUN			
44212	1415	semi-retired	ADJ			
44213	1414	malnourishment	NOUN			
44214	1414	finger-like	ADJ			
44215	1414	photodynamic	ADJ			
44216	1414	curlew	NOUN			

Figure 3. Words near #12200, 24200, and 44200 in the iWeb word frequency list

Notice the average number of tokens for the words at these different frequency

levels in iWeb. Near #12,200 it is about 37,000 tokens (vs about 1,200 in COCA); near #24,200 it is about 8,100 tokens (vs about 300 in COCA); and near #44,200 it is about 1,400 tokens in iWeb (compared to about 50-60 in COCA).

One might argue that 300 tokens would be adequate, which is what one would have for words near word #24,200 in COCA. But in terms of word-level phenomena like collocates, it really is not sufficient. Consider for example the word *alabaster* (“a fine-grained, translucent form of gypsum, typically white, often carved into ornaments”). This word is #24,285 in COCA, and it occurs 290 times in that corpus (see Figure 2 above). In iWeb *alabaster* is actually a bit lower at word #28,576, and it occurs about 5,300 times in iWeb. Figure 4 comes from the “collocates” display for *alabaster* in iWeb:

+ NOUN		NEW WORD	?
244	4.68	box	
192	7.52	jar	
175	4.58	skin	
172	8.52	marble	
125	5.19	stone	
111	2.61	woman	
107	10.00	ointment	
99	7.05	statue	
80	4.43	kitchen	
77	7.38	tomb	
71	3.96	glass	
64	9.11	cavern	
57	7.67	gloss	
56	3.06	wall	
55	7.58	mosque	

+ ADJ		NEW WORD	?
196	4.46	white	
55	2.96	beautiful	
47	2.64	fine	
34	4.05	pure	
33	3.18	expensive	
32	5.18	precious	
32	7.87	translucent	
30	3.49	smooth	
29	5.55	costly	
26	2.81	soft	
26	8.05	carved	
24	5.58	pale	
24	5.96	Egyptian	
17	3.98	genuine	
16	6.75	monumental	

+ VERB		NEW WORD	?
113	2.55	bring	
54	2.68	break	
46	6.61	carve	
29	3.66	paint	
21	3.80	pour	
15	3.08	house	
15	4.26	depict	
15	11.85	onyx	
13	2.77	mount	
13	5.48	pave	
13	7.10	inscribe	
12	3.87	decorate	
11	5.93	sculpt	
10	3.72	frame	
9	3.10	compose	

+ ADV		NEW WORD	?
9	3.37	beautifully	
6	5.85	richly	
6	6.98	intricately	
5	2.82	BC	
5	3.01	nearby	
5	4.43	finely	
4	6.05	delicately	
2	2.63	amazingly	
2	3.52	distinctly	
2	4.29	flawlessly	
2	4.80	beneath	
2	5.51	predominately	
2	5.85	faintly	
2	8.20	ornately	
1	2.50	morally	

Figure 4. Collocates of *alabaster* in iWeb

As one can see, there is very rich collocational data in iWeb, even for a word that is down near #30,000 in the list. There are 97 different noun collocates of *alabaster* in iWeb that occur 10 times or more. In COCA, however, the data is much more sparse. There are only three noun collocates that occur 10 times or more: *skin, face*, and *marble*.

SKIN	41	
FACE	12	
MARBLE	10	
STATUE	9	
WHITE	8	
HAIR	7	
COLUMNS	6	
HEAD	6	
STONE	6	

Figure 5. Noun collocates of *alabaster* in COCA

In the British National Corpus, there are no noun collocates that occur 10 times or more (or even 5 times or more).

HAND	4
MARBLE	4
TOMB	3
RELIEF	3
BT	2
MONUMENT	2
SKIN	2

Figure 6. Noun collocates of *alabaster* in the BNC

To take just one more mid-frequency example, *panini* ("a grilled sandwich made with Italian bread") is word #31,000 in iWeb. The word occurs about 4,280 times in the corpus, and thus the collocates are quite informative. In total, there are 106 noun collocates that occur 7 times or more in iWeb:

+ NOUN		NEW WORD	?
419	9.75	sandwich	
238	8.92	grill	
209	7.65	maker	
134	7.46	salad	
120	6.57	cheese	
106	7.10	pizza	
98	6.51	bread	
89	5.87	chicken	
70	6.77	soup	
69	4.53	press	
62	7.29	pasta	
49	4.06	recipe	
49	6.88	wrap	

+ ADJ		NEW WORD	?
128	9.75	grilled	
62	3.93	hot	
56	5.52	delicious	
39	5.65	Italian	
35	9.11	toasted	
33	3.51	fresh	
23	5.64	homemade	
21	5.70	tasty	
19	2.77	favorite	
18	6.30	baked	
16	2.83	warm	
16	4.66	medium	
16	6.55	roasted	

+ VERB		NEW WORD	?
677	8.42	press	
114	9.72	grill	
85	3.47	serve	
82	5.43	cook	
47	9.93	sandwich	
39	3.70	order	
39	5.92	heat	
36	2.57	place	
33	2.59	eat	
28	7.86	preheat	
27	2.90	close	
27	8.10	toast	
18	5.56	machine	

+ ADV		NEW WORD	?
16	6.30	freshly	
10	2.61	possibly	
4	2.97	surprisingly	
4	6.16	nonstop	
3	2.91	lightly	
3	3.31	evenly	
2	3.33	without	
2	4.50	plus	
2	6.05	inexplicably	
1	2.73	admittedly	
1	2.83	neatly	
1	2.89	approx	
1	2.90	conversely	

Figure 7. Collocates of *panini* in iWeb

In COCA the word panini only has three nouns (*sandwich, salad*, and *grill*) that occur seven times or more (see Figure 8), and in the BNC there are none.

CONTEXT	FREQ
[SANDWICH]	12
[SALAD]	10
[GRILL]	9
[MAKER]	6
[PRESS]	6
[CHEESE]	5
[CHICKEN]	4

Figure 8. Noun collocates of *panini* in COCA

In this section we have given only two examples of medium-frequency words (*alabaster* and *panini*), but the same holds true for most other medium-frequency (and certainly lower-frequency) words.

There is no magical word frequency limit at which the collocates are frequent enough to be helpful and meaningful. But to give some idea of the robustness of collocates at a given word frequency range, consider the lemmas with a frequency of about 500 tokens in the BNC. These would be (adj) *Portuguese, tedious, split, congressional, dire, contracting, pioneering, incoming, stern, youthful*; (noun) *freezer, crossroads, thinker, charcoal, curse, countess, policy-making, aggregate, nostril, lyrics*; and (verb) *sanction, accustom, bypass, banish, tense, enroll, outweigh, border, augment, unload*. Assume that we want seven tokens of a node word + collocate, e.g. *dire* + *consequence*, *dire* + *need*, or *dire* + *warning*. Looking just at noun collocates, and with a collocational span of 4 words left / 4 words right, there are an average of about 14 collocates that occur at least 7 times with each the ten adjectives listed above (*Portuguese, tedious*⋯), an average of about 8 collocates that occur at least 7 times with the nouns, and an average of about 12 collocates that occur at least 7 times with the verbs.

Again, there is no magical number of times that each collocate should occur with a given node word. But assuming that the results just shown are towards the lower end of what we would want as far as collocates, then we might want at least 500 tokens of a given lemma in order to have this minimal number of collocates. And if so, the following are some words that would not meet this threshold of 500 tokens in the BNC: (adj) *diabetic, organized, aristocratic, brisk, intolerable, triumphant, luxurious, rhythmic, irrational, rewarding, stained, economical, discrete, mystical, perpetual, woolen, lifelong, homogeneous, inexperienced, sentimental*; (noun) *mustard, groom, parlor, clip, camel, parrot, bruise, digestion, homeland, ensemble, carcinoma, shuttle, flute, crane, diver, tub, patio, shutter, cuisine, mint*; (verb) *unload, hamper, gesture, sting, shield, lobby, reclaim, disappoint, lessen, flap, revolve, reap, stray, blind, nick, log, contaminate, enrich, prosper, discern*. For words such as these (some of which will probably seem like fairly common and well-known to native speakers of English), we need much larger corpora, like iWeb.

5. The challenges of very large corpora

To this point we have only discussed the advantages of very large corpora (like iWeb), in the domains of syntax, morphology, lexis, and word meaning and usage (via collocates). But very large corpora also bring their challenges and disadvantages, at least potentially. In this section we will discuss the issue of efficiently searching such large corpora, and well as the issue of "granularity".

5.1 Search speed

In terms of search speed, many large corpora are relatively slow, because there is so much data to search through. For example, Table 4 shows (in the rightmost column) the actual search times in seconds for some strings in the Sketch Engine "English Web 2015" (enTenTen15) corpus, which is about 19 billion words (compared to about 14 billion words for iWeb). Note that Sketch Engine divides the search into two parts – finding all matching concordance lines, and then finding the frequency of the matching strings – and the numbers here represent the total of these two searches.

Table 4. Search times in various corpora

Corpus Corpus size	BNC 100m	COCA 560m	GloWbE 1,900m	iWeb 13,900m	SketchEng 19,000m
better to VERB	1.6	1.9	2.2	2.4	32
even more ADJ	1.7	2.5	2.7	3.1	36
best NOUN	1.6	2.5	4.0	4.3	66
if they VERB	1.8	2.5	3.8	1.1	24
has been _vvn (*has been fixed*)	2.3	3.2	4.1	1.4	15
can PRON VERB	2.0	2.6	4.4	1.1	23
VERB some NOUN	3.4	4.2	4.5	1.1	57
the _jjt NOUN (*the biggest piece*)	3.4	4.2	4.5	1.5	75
the NOUN	12.4	*	NA	1.6	126
ADJ NOUN	6.4	NA	NA	1.7	188

Part of the reason that Sketch Engine is rather slow is because it apparently parses the search string linearly. In other words, a search like *the stretcher is* (65 tokens in enTenTen15) takes about 28 seconds, which is almost as much as the

much more frequent string *the story is* (55,890 tokens; 30 seconds). Apparently, the search algorithm searches for all tokens of *the* (which is of course extremely common), and only after it finds these does it check to see if the following word is *story* or *stretcher*, etc. Therefore, even one high frequency word (especially if it is at the beginning of the search string) creates real problems.

As Table 4 indicates, the BYU corpora are much faster than the Sketch Engine corpora. For example, the search *better to VERB* in iWeb takes about 2.4 seconds, which is about 13 times as fast as enTenTen15, even though enTenTen15 is only about 35% bigger than iWeb. (So taking into account the larger size in Sketch Engine, iWeb is still about 9-10 as fast as Sketch Engine.) Part of this is because of the way that searches are done in the BYU corpora. The search string is first parsed to find the least frequent "slot". For example, in the example of *the story is* and *the stretcher is* (shown above), the search would look for *story* or *stretcher*, and only then does it look for cases whether they are preceded by *the*, which makes the search much faster.

Perhaps even more interesting than the comparison of the BYU corpora and Sketch Engine is the fact that corpus size has much less of an impact on search times with the BYU corpus architecture than it does with other architectures. For example, one might imagine that the two billion word GloWbE corpus might be about 20 times as slow as the 100 million word British National Corpus. But in fact it is typically less than twice as slow. Even more surprising is the comparison of the two billion word GloWbE corpus and the 14 billion word iWeb corpus. In the case of searches like *better to VERB*, *even more ADJ*, or *best NOUN*, iWeb is just about 10% slower than GloWbE, rather than the 700% slower that corpus size alone would suggest. The reason for this is the fact that the BYU corpus architecture is built on top of relational databases (specifically SQL Server). This architecture uses many indexes (especially clustered indexes), which speed things up immensely.

Perhaps the most interesting fact is that for the highest frequency searches (*if they VERB*, *has been _vvn*, *ADJ NOUN*, etc), iWeb is actually faster than COCA (which is 1/25th the size) or the BNC (1/140th the size). This is because iWeb quickly parses the search string "slot by slot" (as described above), and it knows that these strings will be very high frequent. For these searches, rather than searching the main corpus databases, it uses "n-grams tables" with the top 10

million or top 100 million for each of the 2-grams (two word strings), 3-grams, 4-grams, and 5-grams in iWeb. Searching through these n-grams databases is much faster than searching through all 14 billion words of data. In essence, then, there are virtually no searches in iWeb that would take more than 4-5 seconds to search the 14 billion words of data (and most take just 1-2 seconds), whereas in Sketch Engine (or even the other BYU corpora, which don't use n-grams), these queries would take 100-200 times as long, or perhaps even just "time out".

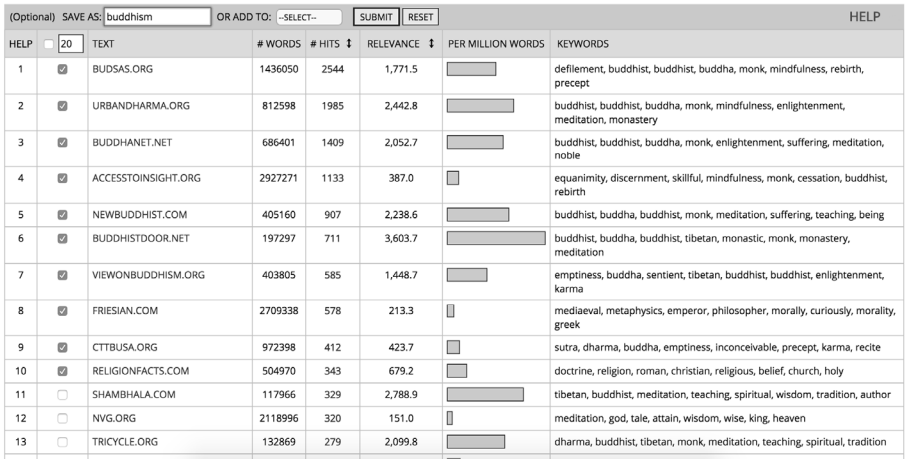
The ability to carry out searches involving high frequency words is not just of theoretical interest. Just as the ability to send spacecraft into orbit around the Earth allowed us to see entire continents at once, the ability to quickly search for virtually anything in the 14 billion words allows us to look at phrases and constructions that would otherwise be too large to investigate. For example, consider the "*way* construction" discussed above: *make your way to, find their way into, worked his way through, navigate your way through*, etc. In iWeb it takes just 2-3 seconds for this search, whereas it is more than 60 seconds in the enTenTen15 corpus. An example of another search with high frequency "slots" is the "quotative like construction" (cf. Tagliamonte and D'Arcy 2004, Buchstaller and D'Arcy 2009, Barbieri 2009), which takes less than two seconds to find all of the strings like *and I was like , 'That has to change'* or *and I was like ,' Well, we just kind of did this and this.'*.

Even higher frequency constructions are very doable, such as *from* ADJ *to* ADJ (e.g. *from bad to worse, from high to low, from good to great*) or VERB POSS NOUN PREP (e.g. *get your hands on, increase your chances of, try your hand at, take my word for*). In each case, the search takes just 1-2 seconds in iWeb (because of its use of n-grams), whereas it would take 2-3 minutes for just one search in enTenTen15 from SketchEngine.

5.2 Blob of data

The other challenge with very large corpora is that they are essentially just a huge "blob" of data, and it is difficult or impossible to restrict the search to only a particular part of the corpus. In iWeb it is very easy to limit the search to a particular part of the corpus. Suppose that users want to create a corpus

dealing with Buddhism, solar energy, basketball, or Harry Potter. In iWeb they simply search for a given word or phrase like *Buddhism*, and in less than one second it will suggest what it thinks are the best sites (using calculations similar to the log likelihood score), such as those shown in Figure 9:



(Optional) SAVE AS: buddhism OR ADD TO: --SELECT-- SUBMIT RESET HELP

HELP	20	TEXT	# WORDS	# HITS	RELEVANCE	PER MILLION WORDS	KEYWORDS
1	☑	BUDSAS.ORG	1436050	2544	1,771.5		defilement, buddhist, buddhist, buddha, monk, mindfulness, rebirth, precept
2	☑	URBANDHARMA.ORG	812598	1985	2,442.8		buddhist, buddhist, buddha, monk, mindfulness, enlightenment, meditation, monastery
3	☑	BUDDHANET.NET	686401	1409	2,052.7		buddhist, buddhist, buddha, monk, enlightenment, suffering, meditation, noble
4	☑	ACCESSTOINSIGHT.ORG	2927271	1133	387.0		equanimity, discernment, skillful, mindfulness, monk, cessation, buddhist, rebirth
5	☑	NEWBUDDHIST.COM	405160	907	2,238.6		buddhist, buddha, buddhist, monk, meditation, suffering, teaching, being
6	☑	BUDDHISTDOOR.NET	197297	711	3,603.7		buddhist, buddha, buddhist, tibetan, monastic, monk, monastery, meditation
7	☑	VIEWONBUDDHISM.ORG	403805	585	1,448.7		emptiness, buddha, sentient, tibetan, buddhist, buddhist, enlightenment, karma
8	☑	FRIESIAN.COM	2709338	578	213.3		mediaeval, metaphysics, emperor, philosopher, morally, curiously, morality, greek
9	☑	CTTBUSA.ORG	972398	412	423.7		sutra, dharma, buddha, emptiness, inconceivable, precept, karma, recite
10	☑	RELIGIONFACTS.COM	504970	343	679.2		doctrine, religion, roman, christian, religious, belief, church, holy
11	☐	SHAMBHALA.COM	117966	329	2,788.9		tibetan, buddhist, meditation, teaching, spiritual, wisdom, tradition, author
12	☐	NVG.ORG	2118996	320	151.0		meditation, god, tale, attain, wisdom, wise, king, heaven
13	☐	TRICYCLE.ORG	132869	279	2,099.8		dharma, buddhist, tibetan, monk, meditation, teaching, spiritual, tradition

Figure 9. Creating Virtual Corpora

Other large corpora have been created by basically just wandering from website to website, gathering random web pages. This means that the vast majority of websites might have just 2-3 pages, which makes searching by website rather meaningless. But in iWeb, the nearly 100,000 websites have any average of 245 web pages and 140,000 words of data, and no website has less than 30 web pages or 10,000 words. So when we search for a topic like Buddhism, solar energy, basketball, or Harry Potter, the matching websites really do deal with this topic. Evidence for this can be found in Figure 9, which shows the most useful keywords for each of the websites in the *Buddhism* search, and similar keyword data is available for all of the nearly 100,000 websites. Users simply select the desired websites from the list shown, and they can select hundreds of different websites, containing tens of millions of words of data. It takes just 5-6 seconds to enter the search terms, browse through the websites, and create a Virtual Corpus on almost any topic. As Figure 10 shows, these Virtual Corpora can be on fairly broad topics (e.g. *Buddhism, basketball, investment,* or *linguistics*), or

very narrow topics (like *dachsund*, *carburetor*, *intentionality*, *iPhone*, or *aquifer*):

MY VIRTUAL CORPORA

HELP		↕	↕	LIST NAME ↕	# WEBSITES ↕	# WORDS ↕	FIND KEYWORDS ◉ SPECIFIC ○ FREQ
1	🗑	🔒		AQUIFER	10	2,263,851	NOUN VERB ADJ ADV
2	🗑	🔒		AUTISM	3	4,233	NOUN VERB ADJ ADV
3	🗑	🔒		BASKETBALL	10	716,536	NOUN VERB ADJ ADV
4	🗑	🔒		BUDDHISM	10	11,055,288	NOUN VERB ADJ ADV
5	🗑	🔒		CANDLE	10	297,424	NOUN VERB ADJ ADV
6	🗑	🔒		CARBURETOR	10	6,976,146	NOUN VERB ADJ ADV
7	🗑	🔒		CARPET	10	540,968	NOUN VERB ADJ ADV
8	🗑	🔒		CHOCOLATE	4	3,067	NOUN VERB ADJ ADV
9	🗑	🔒		CONCRETE	10	3,599,794	NOUN VERB ADJ ADV
10	🗑	🔒		DACHSHUND	10	4,757,280	NOUN VERB ADJ ADV
11	🗑	🔒		DRUMS	10	583,131	NOUN VERB ADJ ADV
12	🗑	🔒		HARRY_POTTER	10	9,526,245	NOUN VERB ADJ ADV
13	🗑	🔒		INTENTIONALITY	10	22,490,188	NOUN VERB ADJ ADV
14	🗑	🔒		INVESTMENT	15	57,186,412	NOUN VERB ADJ ADV
15	🗑	🔒		IPHONE	10	1,304,975	NOUN VERB ADJ ADV
16	🗑	🔒		LINGUISTICS	9	1,673,413	NOUN VERB ADJ ADV

Figure 10. Examples of Virtual Corpora

Once a Virtual Corpora has been created, users can browse the websites in that corpus, add or delete websites, compare the frequency of a word or phrase or grammatical construction in their different Virtual Corpora, and most importantly limit their search to a particular Virtual Corpus (for words, phrases, collocates, or even syntactic constructions). They can also see the keywords from a given Virtual Corpus (Figure 11, for keyword from the Buddhism virtual corpora) and see any of these keywords in context in the particular Virtual Corpus (Figure 12):

BUDDHISM [11,055,288 WORDS, 10 WEBSITES] (0.1% OF TOTAL) NOUN VERB ADJ ADV [ALL CORPORA] SAVE LIST

HELP	ENTRY	WORD (CLICK FOR CONTEXT)	FREQ	# WEBSITES	SPECIFIC FREQ 90 5 WEBSITES	ENTIRE CORPUS	EXPECTED
1		MONASTIC	167	6	7,552.9	28	0.0
2		CONTEMPLATIVE	448	5	2,325.1	244	0.2
3		DEFILEMENT	2784	8	993.4	3,549	2.8
4		UNSATISFACTORINESS	238	6	830.3	363	0.3
5		ANATTA	437	7	741.8	746	0.6
6		DISPASSION	488	5	691.3	894	0.7
7		MEDITATOR	1236	8	651.4	2,403	1.9
8		BUDDHAHOOD	645	9	588.9	1,387	1.1
9		BUDDHIST	3769	10	469.1	10,174	8.0
10		FETTER	575	8	440.8	1,652	1.3
11		TIBETAN	151	8	437.6	437	0.3
12		SAMADHI	1338	8	410.9	4,124	3.3
13		PRECEPT	2499	10	383.1	8,261	6.5
14		EQUANIMITY	1364	8	341.6	5,056	4.0
15		IMPERMANENCE	901	9	331.7	3,440	2.7

Figure 11. Keywords from a Virtual Corpus

FIND SAMPLE: 100 200 500
PAGE: << < 1 / 10 > >>

CLICK FOR MORE CONTEXT [?] SAVE LIST CHOOSE LIST ---------- CREATE NEW LIST [?] SHOW DUPLICATES

1	newbuddhist.com	A	B	C	of change, the thorough-going nature of selflessness. Nagarjuna makes it abundantly clear that impermanence (the relative) is total, complete, thoroughgoing, Absolute. It's
2	budsas.org	A	B	C	form. Because of our accumulated ignorance and wrong view we do not realize the impermanence of citta which falls away as soon as it has arisen and which is succeeded
3	accesstoinsight.org	A	B	C	these painful experiences she had gone through were only tiny drops in the ocean of impermanence in which all beings drown if they are attached to that which rises and cea
4	viewonbuddhism.org	A	B	C	. Impermanence is one of the three Marks of Existence; suffering, non-self and impermanence. Even if we rationally understand these concepts, our view of life only really
5	newbuddhist.com	A	B	C	that I use to recite my Yamantaka mantra is there to REINFORCE my meditation on impermanence... on death. And conquering death. # And it's okay to do
6	buddhanet.net	A	B	C	permanent relationship with anything, at all. # If we examine the notion of impermanence closely and honestly, we see that it is all-pervading, everything is marked by
7	buddhanet.net	A	B	C	" suffering " and " non-self ". From these truths of life, i.e. impermanence, unsatisfactoriness and non-self, how can we establish the significance of our lives?
8	buddhanet.net	A	B	C	have in common. These are the samanna.lakkhana, the universal characteristics of annica (impermanence, changing), dukkha (unsatisfactoriness) and anatta (not-me, not-my
9	urbandharma.org	A	B	C	those of senior monks or lay people, to remind you of the truth of impermanence. Contemplating death is the most profound meditation because it has the power to cut
10	urbandharma.org	A	B	C	seeing this, gave up her pride in her beauty and came to realize the impermanence of beauty. The Buddha, knowing the state of her mind, delivered a
11	accesstoinsight.org	A	B	C	dukkha (unsatisfactoriness) is but a logical corollary arising from this law of universal impermanence. For the impermanent nature of everything can but lead to one inescapal
12	budsas.org	A	B	C	Abhidharma uses the ultimate standpoint. Yet there are passages in the sutras that describe impermanence, impersonality or insubstantiality, elements, and aggregates, and
13	buddhanet.net	A	B	C	this point, which is the high point of a meditator? s sensitivity to impermanence, the sixth purification, purification of knowing and seeing the way, begins.
14	budsas.org	A	B	C	of Sarvodaya is grounded in basic Theravada teachings: the three characteristics of existence (impermanence, suffering, not-self), the mutually interdependent and coarising r
15	budsas.org	A	B	C	matter arising and disappearing at every moment, then they will come to comprehend the impermanence of the processes of lifting the foot, and they will also comprehend t
16	viewonbuddhism.org	A	B	C	- Meditate on purification; this may make your potential clear. - Meditate on impermanence: everything changes, even my bad " I " will change for the better
17	accesstoinsight.org	A	B	C	education, and are normally absorbed by them without difficulty. # The idea of impermanence and of ceaseless change, due to the never-ending " chain " of causes and
18	buddhanet.net	A	B	C	realm, the form realm, and the formless realm) and hence subject to impermanence. When the devas exhaust their celestial blessed-rewards, they will degenerate and fall int

Figure 12. Concordance lines for a keyword from a Virtual Corpus

A corpus with 10-20 billion words can be overwhelming. But the ability to quickly create and use Virtual Corpora makes the corpus much more manageable.

6. Lexically-oriented searches in iWeb

Linguists tend to use corpora to look at things like syntactic and semantic variation. But many language teachers and learners use corpora to look at

detailed information on specific words – in a sense, rather like a "high-powered" dictionary or thesaurus. The iWeb corpus has been designed from the ground up to meet the needs of these users as well.

After creating a clean, accurate list of the top 60,000 words (lemmas) in iWeb, users can easily and quickly browse through that list. Samples of words at #3,600, #23,600, and #43,600 are shown in Figure 13:

	RANK	FREQ	Word	PoS	Audio	Video	Image	KO
1	3601	306841	bid	NOUN				
2	3602	306726	another	PRON				
3	3603	306719	running	NOUN				
4	3604	306438	genetic	ADJ				
5	3605	306149	witness	VERB				
6	3606	306016	plug	NOUN				
7	3607	305960	grocery	NOUN				

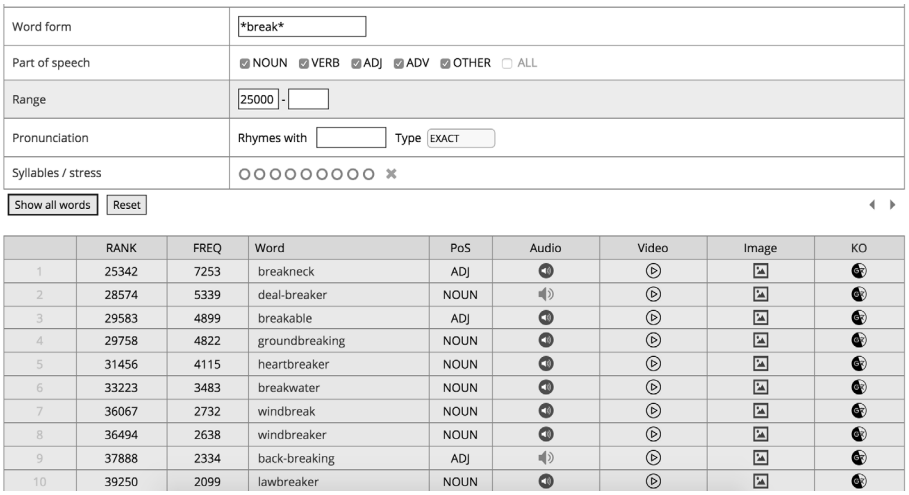
	RANK	FREQ	Word	PoS	Audio	Video	Image	KO
1	23621	8654	denture	NOUN				
2	23622	8653	collectable	ADJ				
3	23623	8648	sordid	ADJ				
4	23624	8647	ascorbic	ADJ				
5	23625	8646	injurious	ADJ				
6	23626	8645	junkyard	NOUN				
7	23627	8644	allay	VERB				

9	43609	1480	augmentative	ADJ				
10	43610	1479	hit-and-miss	ADJ				
11	43611	1479	liquidy	ADJ				
12	43612	1479	biff	NOUN				
13	43613	1479	belligerent	NOUN				
14	43614	1479	unreached	ADJ				
15	43615	1479	spliff	NOUN				

Figure 13. Frequency lists from iWeb; near words #3600, 23600, 43600

For each word, users can hear the pronunciation, find the word used in videos, see a picture of the concept (from Google Images), find a translation into the language of their choice, and of course click on a word for much more detailed information, as is discussed below.

Of course users can also search through this 60,000 word list as well. For example, Figure 14 shows a basic search for words with *break*, starting at word 25,000 in the 60,000 word list.



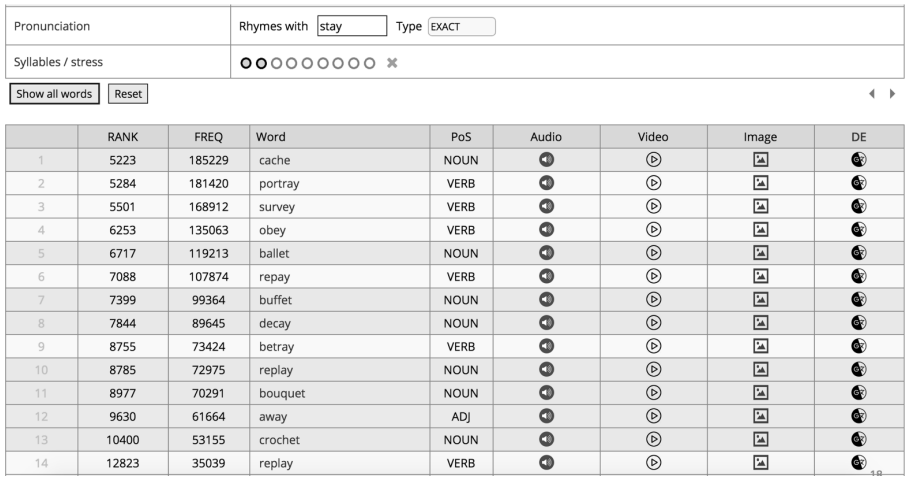
Word form	*break*
Part of speech	NOUN VERB ADJ ADV OTHER ALL
Range	25000 -
Pronunciation	Rhymes with Type EXACT
Syllables / stress	

Show all words Reset

	RANK	FREQ	Word	PoS	Audio	Video	Image	KO
1	25342	7253	breakneck	ADJ				
2	28574	5339	deal-breaker	NOUN				
3	29583	4899	breakable	ADJ				
4	29758	4822	groundbreaking	NOUN				
5	31456	4115	heartbreaker	NOUN				
6	33223	3483	breakwater	NOUN				
7	36067	2732	windbreak	NOUN				
8	36494	2638	windbreaker	NOUN				
9	37888	2334	back-breaking	ADJ				
10	39250	2099	lawbreaker	NOUN				

Figure 14. Searching through the top 60,000 words

Users can even search by pronunciation, which is very useful, considering the notorious difficulty of English orthography. For example, Figure 15 shows the results of a search for two syllable words that are accented on the second syllable, which rhyme with *stay* (notice the spellings in –ay, -ey, -e, and -et:



Pronunciation	Rhymes with stay Type EXACT
Syllables / stress	

Show all words Reset

	RANK	FREQ	Word	PoS	Audio	Video	Image	DE
1	5223	185229	cache	NOUN				
2	5284	181420	portray	VERB				
3	5501	168912	survey	VERB				
4	6253	135063	obey	VERB				
5	6717	119213	ballet	NOUN				
6	7088	107874	repay	VERB				
7	7399	99364	buffet	NOUN				
8	7844	89645	decay	NOUN				
9	8755	73424	betray	VERB				
10	8785	72975	replay	NOUN				
11	8977	70291	bouquet	NOUN				
12	9630	61664	away	ADJ				
13	10400	53155	crochet	NOUN				
14	12823	35039	replay	VERB				

Figure 15. Searching by pronunciation

In terms of information on words, perhaps the most useful feature is the ability to see a wealth of information on each of the top 60,000 words in the corpus. Because each of these word-level pages have already been created ahead of time, each of them can by displayed in one second or less.

Each of the 60,000 words has a "home page" that shows the frequency, definition, links to pronunciation, images, videos, and translations, as well as collocates, related topics, word clusters, websites, and concordance lines. For each of these features, users can click to see a more complete display, as will be shown below. In other words, the "home page" is basically a summary of the top results from each of these more complete pages.

The first of these more detailed pages is the "dictionary" page, which duplicates some of the information from the word "home page", including provides definitions and links to pronunciation, images, videos, and translations. But it also includes detailed frequency information (including range across the corpus), word forms and their frequency, related words, synonyms, and WordNet entries (hypernyms and hyponyms), as in the entry for *stream* in Figure 16.

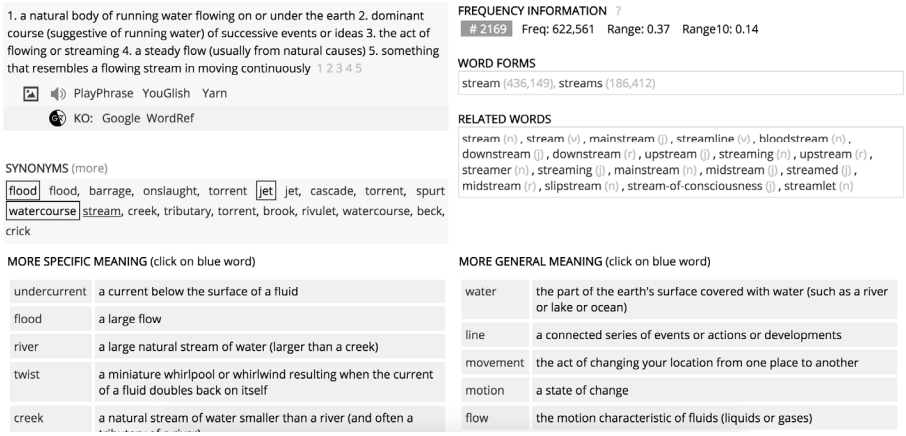


Figure 16. "Dictionary" page (for *stream*)

There is also a "collocates" page for each of the top 60,000 words (as in Figure 17, for *bread*), which groups the collocates by part of speech, and which shows the most frequent position of the collocate with regards to the node word, the

Mutual Information score, a link to see the node word and collocate in context, and links to see a new collocates page for any of the linked collocates (in this way, users can browse through a network of related words). Users can also sort by Mutual Information and limit the collocates by frequency (via “Advanced Options”).

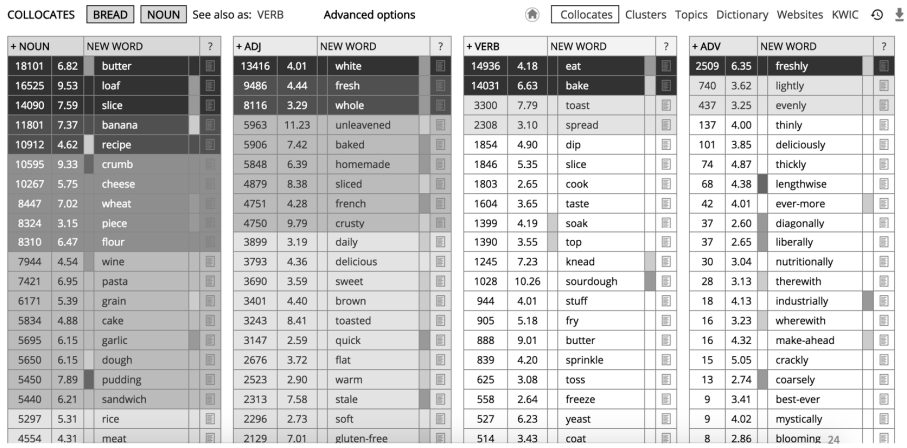
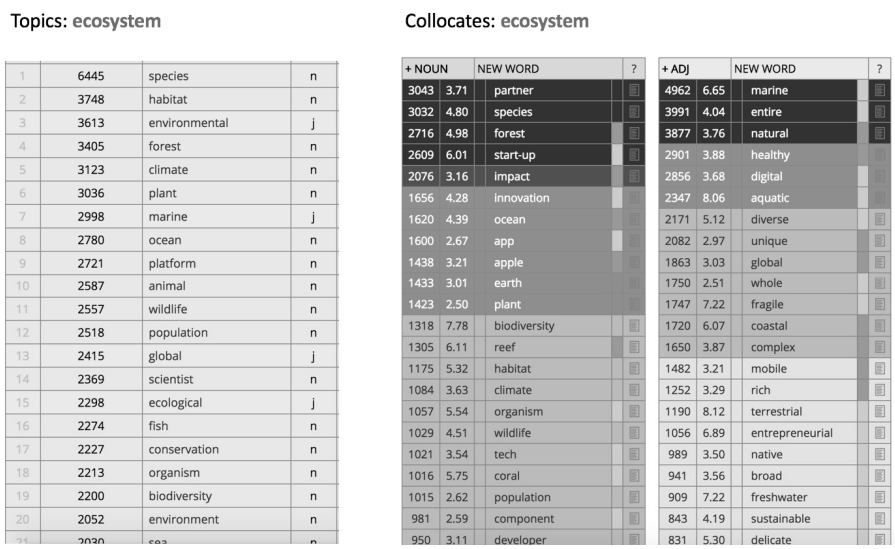


Figure 17. “Collocates” page (for *bread*)

The collocates page shows nearby words (typically within a span of 4 words left and 4 words right). But a separate “topics” page shows the co-occurring words *anywhere* on the web page. Figure 18 (for *ecosystem*) compares the topics display and the collocates display. As with collocates, users can click on any of the related topics, and thus explore a “chain” of semantically-related concepts.



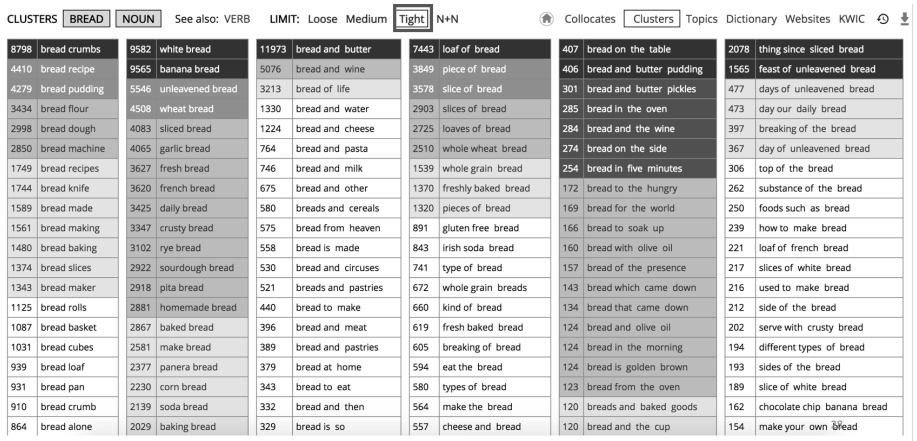
1	6445	species	n
2	3748	habitat	n
3	3613	environmental	j
4	3405	forest	n
5	3123	climate	n
6	3036	plant	n
7	2998	marine	j
8	2780	ocean	n
9	2721	platform	n
10	2587	animal	n
11	2557	wildlife	n
12	2518	population	n
13	2415	global	j
14	2369	scientist	n
15	2298	ecological	j
16	2274	fish	n
17	2227	conservation	n
18	2213	organism	n
19	2200	biodiversity	n
20	2052	environment	n
21	2030	sea	n

+ NOUN		NEW WORD
3043	3.71	partner
3032	4.80	species
2716	4.98	forest
2609	6.01	start-up
2076	3.16	impact
1656	4.28	innovation
1620	4.39	ocean
1600	2.67	app
1438	3.21	apple
1433	3.01	earth
1423	2.50	plant
1318	7.78	biodiversity
1305	6.11	reef
1175	5.32	habitat
1084	3.63	climate
1057	5.54	organism
1029	4.51	wildlife
1021	3.54	tech
1016	5.75	coral
1015	2.62	population
981	2.59	component
950	3.11	developer

+ ADJ		NEW WORD
4962	6.65	marine
3991	4.04	entire
3877	3.76	natural
2901	3.88	healthy
2856	3.68	digital
2347	8.06	aquatic
2171	5.12	diverse
2082	2.97	unique
1863	3.03	global
1750	2.51	whole
1747	7.22	fragile
1720	6.07	coastal
1650	3.87	complex
1482	3.21	mobile
1252	3.29	rich
1190	8.12	terrestrial
1056	6.89	entrepreneurial
989	3.50	native
941	3.56	broad
909	7.22	freshwater
843	4.19	sustainable
831	5.30	delicate

Figure 18. "Related topics" page (for *ecosystem*)

The "word clusters" page shows the most frequent 2, 3, and 4-word strings for a given word, as in Figure 19 (for *bread*). Users can also choose how "tight" to make the clusters, as far as eliminating or including strings with high frequency words (Figure 19 is set to "tight clusters").



8798	bread crumbs
4410	bread recipe
4279	bread pudding
3434	bread flour
2998	bread dough
2850	bread machine
1749	bread recipes
1744	bread knife
1589	bread made
1561	bread making
1480	bread baking
1374	bread slices
1343	bread maker
1125	bread rolls
1087	bread basket
1031	bread cubes
939	bread loaf
931	bread pan
910	bread crumb
864	bread alone

9582	white bread
9565	banana bread
5546	unleavened bread
4508	wheat bread
4083	sliced bread
4065	garlic bread
3627	fresh bread
3620	french bread
3425	daily bread
3347	crusty bread
3102	rye bread
2922	sourdough bread
2918	pita bread
2881	homemade bread
2867	baked bread
2581	make bread
2377	panera bread
2230	corn bread
2139	soda bread
2029	baking bread

11973	bread and butter
5076	bread and wine
3213	bread of life
1330	bread and water
1224	bread and cheese
764	bread and pasta
746	bread and milk
675	bread and other
580	breads and cereals
575	bread from heaven
558	bread is made
530	bread and circuses
521	breads and pastries
440	bread to make
396	bread and meat
389	bread and pastries
379	bread at home
343	bread to eat
332	bread and then
329	bread is so

7443	loaf of bread
3849	piece of bread
3578	slice of bread
2903	slices of bread
2725	loaves of bread
2510	whole wheat bread
1539	whole grain bread
1370	freshly baked bread
1320	pieces of bread
891	gluten free bread
843	irish soda bread
741	type of bread
672	whole grain breads
660	kind of bread
619	fresh baked bread
605	breaking of bread
594	eat the bread
580	types of bread
564	make the bread
557	cheese and bread

407	bread on the table
406	bread and butter pudding
301	bread and butter pickles
285	bread in the oven
284	bread and the wine
274	bread on the side
254	bread in five minutes
172	bread to the hungry
169	bread for the world
166	bread to soak up
160	bread with olive oil
157	bread of the presence
143	bread which came down
134	bread that came down
124	bread and olive oil
124	bread in the morning
124	bread is golden brown
123	bread from the oven
120	breads and baked goods
120	bread and the cup

2078	thing since sliced bread
1565	feast of unleavened bread
477	days of unleavened bread
473	day our daily bread
397	breaking of the bread
367	day of unleavened bread
306	top of the bread
262	substance of the bread
250	foods such as bread
239	how to make bread
221	loaf of french bread
217	slices of white bread
216	used to make bread
212	side of the bread
202	serve with crusty bread
194	different types of bread
193	sides of the bread
189	slice of white bread
162	chocolate chip banana bread
154	make your own bread

Figure 19. "Word clusters" page (for *bread*)

Another page allows users to see re-sortable concordance lines for any of the top 60,000 words, as in Figure 20. (This sample page is for *fathom* and is sorted by words to the left, which shows that the word is nearly always preceded by negation).

156	inshape.com	weight on overnight . Finally by Senior year I	could	n't	fathom the way I felt , looked , or acted , it was
157	freetopbooks.com	to leave the place , but for reasons I	could	n't	fathom , I did n't . I stood rooted to that spot for
158	whole-dog-journal.com	adult Golden-Labrador mix who adored car trips . I	could	n't	fathom what had gotten into her -- until a few minutes later when
159	welltrainedmind.com	I know that 's awful ! But I just	could	n't	fathom tipping $100 . # If you have checked out of the room
160	exemplore.com	unnerved to learn hers was a deer ! She	could	n't	fathom a deer having a place as a power animal . Great hub
161	mythreecents.com	charged in season was $163.00 + tax . They	could	n't	fathom that anyone would charge that much money . Clearly they implied
162	www.cricket.com.au	splice fly into space , and for bowlers who	could	n't	fathom why they failed to find the woodwork or the waiting hands of
163	newlovetimes.com	are too picky for their own good . I	can	never	fathom what being picky means ! Okay , so we do n't want
164	thedivinemercy.org	will . These are mysteries that the human mind	will	never	fathom here on earth ; eternity will reveal them (1656) .
165	the-chicken-chick.com	dry . # Purported Use #2 : Insecticide We	would	never	fathom taking antibiotics daily as a preventative due to fear of one day
166	ironworksforum.com	over Dresden at angels one five , Though they	'll	never	fathom it behind my sarcasm desperate memories lie # Will I ever be
168	spartanavenue.com	that theyre more afraid of MSU than ever and	can	not	fathom the idea that Mark Dantonio is the best coach in the state
169	financegourmet.com	to cut out dining out for lunch , but	can	not	fathom doing anything to lower their car payment . # That $800 in
170	greetingcardpoet.com	, and in the recognition that your own children	can	not	fathom the depth of your love you come to understand the tragic ,
171	citronresearch.com	- Citron reviews IBOCs loans of record # Citron	can	not	fathom how IBOC expects investors to believe its current bad loan
172	www.counterpunch.org	In spite of the fact that this Trump demographic	can	not	fathom what the hell is going on with our globalized , financialized
173	famguardian.org	shall not cease to believe in them because I	can	not	fathom them , and I had rather mistrust my own capacity than his
174	thefinancialdiet.com	DiCaprio sighting is just too fucking much . I	can	not	fathom the last time I spent $400 on one evening , but I

Figure 20. "Keyword in Context" page (for *fathom*)

Finally, users can find the top websites for each of the top 60,000 words in the corpus, as in Figure 21 (for *coffee*). Note that iWeb already knows the top keywords for each website, which helps users to get a "snapshot" picture of that website.

Create Virtual Corpus

Freq	Per 1000	% Pages		Website (click for topics)	Words
2707	44.3	90%		coffee.org	coffee, brew, brewing, bean, maker, tea, pot, cup, filter, flavor, machine, taste, hot, water, perfect, variety, type, serve, different, home,
1328	43.1	98%		coffeeresearch.org	espresso, coffee, roast, bean, flavor, grow, size, produce, quality, often, should, high, information, best, find, more, then, other, only,
2910	38.1	99%		driftaway.coffee	coffee, roast, brew, roast, brew, brewing, bean, flavor, cup, tag, region, process, ground, profile, produce, method, water, grow, different, might,
2233	35.9	100%		roastycoffee.com	coffee, brew, grinder, brew, brewing, bean, grind, drip, maker, cup, pour, pot, taste, filter, taste, drink, heat, lover, flavor, kitchen,
2606	34.1	73%		coffeeforless.com	coffee, brew, beverage, flavor, cup, taste, enjoy, quality, office, buy, home, service,
2581	33.4	93%		coffee-brewing-methods.com	espresso, brewing, coffee, brew, drip, extraction, bean, maker, french, filter, cup, taste, machine, temperature, ground, hot, method, water, device, prepare,
880	32.1	93%		1912pike.com	barista, espresso, brew, coffee, bean, flavor, cup, rich,

Figure 21. "Websites" page (for *coffee*)

Users can then click on “Create Virtual Corpus”, and within one second they have a Virtual Corpus with millions of words on a particular topic. As explained above, they can then limit their searches (words, phrases, collocates) to this Virtual Corpus, as well as compare features in their different Virtual Corpora.

(Optional) SAVE AS: coffee OR ADD TO: --SELECT-- SUBMIT RESET HELP

HELP	20	TEXT	# WORDS	# HITS	RELEVANCE	PER MILLION WORDS	KEYWORDS
1	☑	COFFEE.ORG	61075	2718	44,502.7		coffee, brew, brewing, bean, maker, tea, pot, cup
2	☑	COFFEERESEARCH.ORG	30783	1328	43,140.7		espresso, coffee, roast, bean, flavor, grow, size, produce
3	☑	DRIFTAWAY.COFFEE	76448	2928	38,300.5		coffee, roast, brew, roast, brew, brewing, bean, flavor
4	☑	COFFEEFORLESS.COM	76475	2772	36,247.1		coffee, brew, beverage, flavor, cup, taste, enjoy, quality
5	☑	CREMA.CO	33481	1213	36,229.5		roaster, acidity, coffee, volcanic, grower, cherry, farmer, specialty
6	☑	ROASTYCOFFEE.COM	62180	2241	36,040.5		coffee, brew, grinder, brew, brewing, bean, grind, drip
7	☑	COFFEE-BREWING-METHODS.COM	77238	2639	34,167.1		espresso, brewing, coffee, brew, drip, extraction, bean, maker
8	☑	TODDYCAFE.COM	20372	654	32,102.9		toddy, iced, brew, brewing, coffee, brew, concentrate, bean
9	☑	1912PIKE.COM	27433	880	32,078.2		barista, espresso, brew, coffee, bean, flavor, cup, rich
10	☑	HAWAIICOFFEECOMPANY.COM	32785	1038	31,660.8		coffee, lion, flavor, con, blend, rating, flavor, pro
11	☐	COFFEECHEMISTRY.COM	62725	1885	30,051.8		caffeine, coffee, roast, bean, acid, compound, quality, level
12	☐	FRESHROASTEDCOFFEE.COM	56462	1694	30,002.5		roast, roasted, coffee, con, bean, tea, blend, pro

Figure 22. Virtual Corpus (for *coffee*)

7. Conclusion

New technologies have allowed corpus creators to create very large, multi-billion word corpora from texts on the web. These large corpora allow researchers to examine a wide range of syntactic, morphological, lexical, and semantic phenomena in ways that would be quite impossible with a much smaller 100 million or 500 million word corpus. The challenge, however, is to not be overwhelmed with these immense, “blobs” of data. With the right corpus architecture, users can search through 14 billion words of data (as in iWeb) in not much more time that it would take to search a corpus less than 1/100th that size (as with the British National Corpus). In addition, users can quickly and easily create Virtual Corpora for words and topics of interest, and then search just with those Virtual Corpora, and compare across them.

Many corpora are oriented more towards linguists, with the ability to search by word, lemma, and part of speech. The iWeb corpus allows all of these, and more. For example, the search in brackets { *VERB* * =*EXPENSIVE* @*clothes* } would find any verb + any word + any form (hence the caps) of any synonym of

expensive + any word in a customized "clothes" wordlist that they have created (e.g. *bought some expensive shoes, wearing some costly shoes*). And because of the advanced corpus architecture, even a complex query like this would take just 1-2 seconds to search the entire 14 billion words in the corpus.

But iWeb is also designed for learners and teachers – not just linguists. In this sense, perhaps the most useful features are the ability to search through the top 60,000 words (by word form, part of speech, word frequency, and even pronunciation) and then to see a wealth of information on each of these words – including frequency information, definition, links to pronunciation, images, and videos; word forms, related words, and synonyms; and collocates, related topics, word clusters, concordance lines, and related websites.

In summary, the right corpus architecture and interface allows users to easily access extremely large amounts of data, in ways that would have been unthinkable 10-15 years ago – all of which can help to transform teaching, learning, and research.

References

Barbieri, Federica. 2009. Quotative 'be like' in American English: Ephemeral or here to stay? *English World-Wide* 30: 68-90.

Buchstaller, Isabelle and Alex D'Arcy. 2009. Localized globalization: A multi-local, multivariate investigation of quotative 'be like'. *Journal of Sociolinguistics* 13: 291-331.

Davies, Mark and Jong-Bok Kim. 2019. Historical shifts with the into-causative construction in American English. *Linguistics* 57: 29-58.

Goldberg, Adele. 1997. Making one's way through the data. In Masayoshi Shibatani and Sandra A. Thompson (eds.), *Grammatical constructions: Their form and meaning*, 29-53. Oxford: Clarendon Press.

Israel, Michael. 1996. The way constructions grow. In Adele Goldberg (ed.), *Conceptual structure, discourse and language*, 217-230. Stanford, CA: CSLI.

Kim, Jong-Bok and Mark Davies. 2016. The into-causative construction in English: A construction-based perspective. *English Language and Linguistics* 20: 55-83.

Mair, Christian. 2002. Three changing patterns of verb complementation in late modern English: A real-time study based on matching text corpora. *English Language and Linguistics* 6: 106-131.

Rohdenburg, Gunter. 2007. Functional constraints in syntactic change: The rise and fall of

prepositional constructions in early and late modern English. *English Studies* 88: 217-233.

Rudanko, Juhani. 2000. *Corpora and complementation: Tracing sentential complementation patterns of nouns, adjectives, and verbs over the last three centuries*. Lanham, MD: University Press of America.

Rudanko, Juhani. 2012. Exploring aspects of the Great Complement Shift, with evidence from the TIME Corpus and COCA. In Terttu Nevalainen and Elizabeth Closs Traugott (eds.), 222-232. *The Oxford Handbook of the History of English.* New York, NY: Oxford University Press.

Stange, Ulrike. 2017. You're so not going to believe this: The use of genx so in constructions with future going to in American English. *American Speech* 92: 487-524.

Tagliamonte, Sali and Alex D'Arcy. 2004. He's like, she's like: The quotative system in Canadian Youth. *Journal of Sociolinguistics* 8: 493-514.

Vosberg, Uwe. 2003. Cognitive complexity and the establishment of *-ing* constructions with retrospective verbs in modern English. In Marina Dossena and Charles Jones (eds.), *Insights into Late Modern English*, 197-220. Bern: Peter Lang.

Mark Davies
Professor
Linguistics
Brigham Young University
Provo, UT 84602, U.S.A
E-mail: mark_davies@byu.edu

Jong-Bok Kim
Professor
Dept. of English Language and Literature
Kyung Hee University
26, Kyungheedae-ro, Dongdaemun-gu,
Seoul, 02447, Rep. of Korea
E-mail: jongbok@khu.ac.kr

Received: 2019. 01. 07.
Revised: 2019. 03. 01.
Accepted: 2019. 03. 09.