Rocking the BYU COCA: Uses for Research and Pedagogy

Corpus of Contemporary American English – Workshop November 2015

BYU COCA Texts: Registers and sources

The corpus is composed of more than 450 million words in 189,431 texts, including 20 million words each year from 1990-2012. **The most recent addition of texts (Apr 2011 - Jun 2012) was completed in June 2012.** Because of its design, it is **perhaps the only corpus of English that is suitable for looking at current, ongoing changes in the language** (see the 2011 article in Literary and Linguistic Computing).

For each year (and therefore overall, as well), the corpus is evenly divided between the five genres of spoken, fiction, popular magazines, newspapers, and academic journals. The texts come from a variety of sources:

- **Spoken**: (95 million words [95,385,672]) Transcripts of unscripted conversation from more than 150 different TV and radio programs (examples: *All Things Considered* (NPR), *Newshour* (PBS), *Good Morning America* (ABC), *Today Show* (NBC), *60 Minutes* (CBS), *Hannity and Colmes* (Fox), *Jerry Springer*, etc). [See notes on the naturalness and authenticity of the language from these transcripts).
- **Fiction**: (90 million words [90,344,134]) Short stories and plays from literary magazines, children's magazines, popular magazines, first chapters of first edition books 1990-present, and movie scripts.
- **Popular Magazines:** (95 million words [95,564,706]) Nearly 100 different magazines, with a good mix (overall, and by year) between specific domains (news, health, home and gardening, women, financial, religion, sports, etc). A few examples are *Time, Men's Health, Good Housekeeping, Cosmopolitan, Fortune, Christian Century, Sports Illustrated*, etc.
- **Newspapers:** (92 million words [91,680,966]) Ten newspapers from across the US, including: *USA Today, New York Times, Atlanta Journal Constitution, San Francisco Chronicle*, etc. In most cases, there is a good mix between different sections of the newspaper, such as local news, opinion, sports, financial, etc.
- **Academic Journals**: (91 million words [91,044,778]) Nearly 100 different peer-reviewed journals. These were selected to cover the entire range of the Library of Congress classification system (e.g. a certain percentage from B (philosophy, psychology, religion), D (world history), K (education), T (technology), etc.), both overall and by number of words per year

COCA and Other BYU Corpora

Click here:
http://corpus.byu.edu/

Other corpora hosted on the BYU website are:

English	# words	language/dialect	time period	compare
NEW Hansard Corpus (British Parliament)	1.6 **billion**	British	1803-2005	Info
Wikipedia Corpus (with virtual corpora)	1.9 **billion**	English	-2014	Info
Global Web-Based English (GloWbE)	1.9 **billion**	20 countries	2012-13	
Corpus of Contemporary American English (COCA)	450 million	American	1990-2012	* * * * *
Corpus of Historical American English (COHA)	400 million	American	1810-2009	* *
TIME Magazine Corpus	100 million	American	1923-2006	
Corpus of American Soap Operas	100 million	American	2001-2012	*
British National Corpus (BYU-BNC)*	100 million	British	1980s-1993	* *
Strathy Corpus (Canada)	50 million	Canadian	1970s-2000s	
Other languages				
Corpus del Español (see also…)	100 million	Spanish	1200s-1900s	*
Corpus do Português (see also…)	45 million	Portuguese	1300s-1900s	
N-grams				
Google Books: American English	155 **billion**	American	1500s-2000s	*
Google Books: British English	34 billion	British	1500s-2000s	
Google Books: One Million Books	89 billion	Am/Br	1500s-2000s	
Google Books: Spanish	45 billion	Spanish	1500s-2000s	

They can actually be used to compare between the corpora. There is a COMPARE function in the display area.

GloWbe:

- Compare across dialects

The Corpus of Global Web-Based English (GloWbE) is composed of 1.9 billion words from 1.8 million web pages in 20 different English-speaking countries.

DIALECTS: The real power of GloWbE, though, is the ability to see the frequency of any word, phrase, or grammatical construction in each of the 20 different countries. You can also compare any features in two sets of dialects, such as British and American English (in more than 775 million words of text for just these two dialects). Or you could just limit your search to one or two countries (e.g. Australia (148 million words), South Africa (45 million), or Singapore (43 million)), and you'll still be searching the largest online corpus for most of these twenty countries.

Wikipedia Corpus:

- Create a personalized virtual corpora

Most importantly, however, with this interface you can quickly and easily create and then search personalized "virtual corpora" from the 4,400,000 web pages. For example, in less than a minute you could create a corpus with 500-1,000 pages (perhaps 500,000-1,000,000 words) related to microbiology, economics, basketball, Buddhism, or thousands of other topics. You can then modify any of these corpora -- adding, deleting, or moving texts

Hansard Corpus:

- <u>Tracking British political discussion</u>

contains nearly every speech given in the <u>British Parliament</u> from 1803-2005, and it allows you to search these speeches (including <u>semantically-based searches</u>) in ways that are not possible with any other resource. Example:

SEE FREQUENCY OF A WORD OR PHRASE OVER TIME

- <u>Increase</u>: (noun) environment, technology, tourism, entitlement (verb) recycle, highlight, update, focus (adjective) regional, terrorist, ethnic, innovative, back-bench (adverb) tragically, forcefully, dramatically, worldwide

- <u>Decrease</u>: (noun) evil, clergyman, admiral, missionary, exertion, zeal (verb) avow, induce, acquiesce, assail (adjective) sacred, cordial, manly, impolitic, landed (adverb) lately, effectually, advantageously

- <u>Increase/decrease ("spike" in frequency)</u>: (noun) colonist (1850s), production (1940s), inflation (1970s), housewife (1940s-1970s), reactor (1970s), motorway (1970s); (verb) entreat (1830s), nationalise (1940s), overspend (1980s); (adjective) industrious (1830s-840s), nuclear (1960s-1980s), self-employed (1970s); (adverb) injuriously (after mid-1800s), appallingly (through 1980s)

SEARCH BOX

A. How many instances of the word "chip" are in the whole corpus?

Activity 1. LIST search. Searching the whole corpus.

➢ Click DISPLAY>LIST
➢ Put in *chip* in the word (s) box.
➢ Click Search. What does the result show?
  o FREQ:

B. What are the most used adjectives which describe the word "chip"?

➢ Click DISPLAY>LIST
➢ Put *chip* in the search box.
➢ Choose "adj. All" from the POS List. Copy the tag **[j*]** in front of the word *chip*
➢ Click "Search"

C. What is the most frequently used adjective in COCA?

➢ Click DISPLAY>LIST
➢ Put **[j*]** in the search box. Click Search.

TAGS: You can always generate the tag symbol using the POS list.

D. List & Chart > What is the difference?

Activity 2. CHART function

➢ Click DISPLAY>LIST
➢ Put in *[nn*] chip* in the word (s) box.
➢ Click Search. What does the result frame show?
  o FREQ:
➢ Switch to Chart. What does the result show?
  o What can we learn from List compared to Chart?

Activity 3. Registers in CHARTS. CHARTS results.

E. Which register uses past tense verbs most frequently (eg. Was, had, were, said, did…? Tag: [v?d*]

➢ Click DISPLAY>CHART
➢ Insert *[v?d*]* in the search field. Click Search.

F. Which subregister in ACADEMIC group uses past tense verbs most frequently (eg. Was, had, were, said, did…?
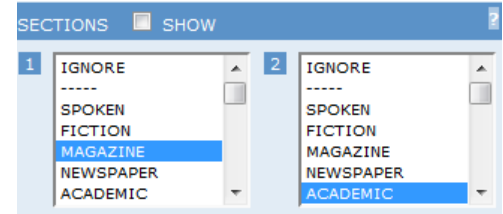
➢   Click on the ACADEMIC column heading.
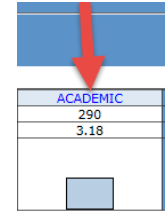
ACADEMIC
290
3.18

Activity 4. Simple collocates comparison across registers.

G. What collocates with the word *hard* in Magazines and Academic registers?

➢   Click DISPLAY>LIST
➢   Insert **hard [nn*]**
➢   From Sections select Magazine on the left and Academic on the right
➢   Click "Search"

Let's analyze the results chart

Activity 5. Advanced collocate search. LIST. Sorting Frequency and Relevance.

G. What is the difference of results between FREQUENCY and RELEVANCE sorting of collocates?

➢   Click DISPLAY>LIST
➢   Input **laugh.[n*]** in the search box
➢   Click Collocates function. Add a * in the field. Choose 5 on the left and 5 on the right.

SEARCH STRING

WORD(S) laugh.[n*]

COLLOCATES *   5 ▾ 5 ▾

POS LIST

RANDOM   SEARCH   RESET
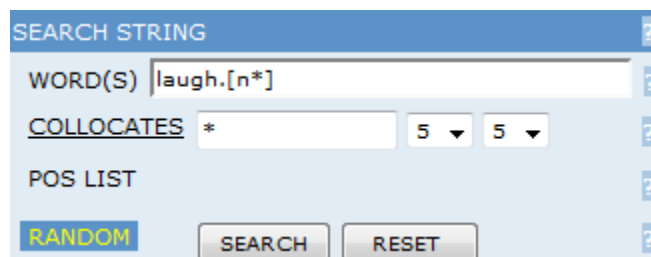
➢   From the SORTING AND LIMITS box choose: Frequency.
  o   Examine the results.
➢   From the SORTING AND LIMITS box choose: Relevance.
  o   Examine the results.

What is the difference?

Use cheat sheet to input any queries you want ☺

Activity 6. KWIC – Keyword in Context

➢ Click DISPLAY>KWIC
➢ Put *research.* in the search field
➢ Click Search

Resorting: This menu can be used before and after the search. Choose these options in the Display/Sorting section of the Search box.

It can also be found in the upper right corner of the KWIC frame.

| L | - | - | - | - | 1 | 2 | 3 | R | | * | | RE-SORT | ? |
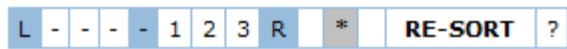
Resorting:
o Click L and "Re-sort"
o Click R and "Re-sort"
o Click the dashes – you will add numbers.

Activity 7. COMPARISON Search

➢ Click DISPLAY>COMPARISON
➢ Put *Democrats* in the first search box, *Republicans* in the second
➢ Click on COLLOCATES. Type [j*] for an adjective in the box.
➢ From SORTING choose Relevance.
➢ Choose 5 for the left node, 5 in the right.

PEDAGOGICAL USES OF BYU CORPORA

Words and Phrase Analysis

http://www.wordandphrase.info

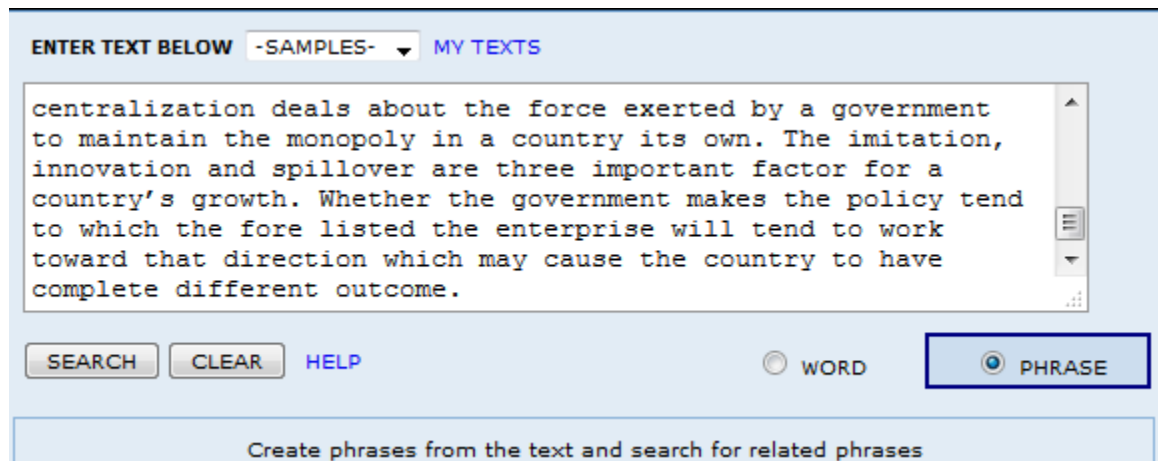This interactive interface of BYU corpora can help our students to efficiently look for best *collocates* for their words.

We are going to use the **Input/analyze text** interface. **(1)** Click and copy and paste this fragment:

SAMPLE TEXT from ENGL106i

Certain country have leaders who strength a lot in the form of development and want to intervene the country's economy in methods for instance Singapore. The technology development can be a great push on the country's economic growth. If you dismember the parts of government influence, economic growth and technology development it's easier to seek for a pattern in it. The role of government can be categorized into two, one is political and one is economical. Economically, government should control the production and usage of resource. Politically government should focus on the behavior of a government. The centralization of the government in economic can be understand as the level of government control the market. The World Bank stresses each country to have an open market instead of a highly controlled market. Political centralization deals about the force exerted by a government to maintain the monopoly in a country its own. The imitation, innovation and spillover are three important factor for a country's growth. Whether the government makes the policy tend to which the fore listed the enterprise will tend to work toward that direction which may cause the country to have complete different outcome.

(2) Choose "**Phrase**" setting:

(3) Click **Search.**

SEE LISTS	FREQ RANGE	1-500	501-3000	> 3000		ACAD	HELP
	194 WORDS	75 %	15 %	9 %		15 %	

Certain country have leaders who strength a lot in the form of development and want to **intervene** the country's economy in methods for instance Singapore. The technology development can be a great push on the country's economic growth. If you **dismember** the parts of government influence, economic growth and technology development it's easier to seek for a pattern in it. The role of government can be **categorized** into two, one is political and one is **economical**. **Economically**, government should control the production and **usage** of resource. **Politically** government should focus on the behavior of a government. The **centralization** of the government in economic can be understand as the level of government control the market. The World Bank stresses each country to have an open market instead of a highly controlled market. Political **centralization** deals about the force **exerted** by a government to maintain the **monopoly** in a country its own. The **imitation**, **innovation** and **spillover** are three important factor for a country's growth. Whether the government makes the policy tend to which the **fore** listed the **enterprise** will tend to work toward that direction which may cause the country to have complete different outcome.

(4) **Click** on the yellow word "intervene"

In the frame below you will see:

- o DEFINITION
- o COLLOCATES (noun, misc)

INTERVENE v (RANK 5144, FREQ 5109)   (BAD

	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
CLICK BAR TO LIMIT					
STORED	46	20	33	36	67
MORE	1107	520	772	880	1830

DEFINITIONS (WORDNET)

1. get involved, usually so as to hinder or halt an action

COLLOCATES (click to see with INTERVENE)

noun government, court, force, affair, behalf, politics, decision, conflict, police, authority, military, matter, dispute, fate, troop **misc** when, directly, militarily, behalf, stop, preven military, unless, early, effectively, refuse, save, internal, force

Below that information you will see Key word in context:

another species to extinction , c'est la guerre . We only	intervene	if a whole	ecosystem is threatened with collapse . That way the
' So the US Supreme Court would not be likely to	intervene	if the state	court was merely interpreting law particularly
force . The official said American soldiers are permitted to	intervene	if they see	' Rodney King-like ' violence . They can take
tell policy . Today , the U.S. Supreme Court refused to	intervene	in a California	case that would have blocked enforcement on the
director of member services . Awtrey said he was prepared to	intervene	in a case	no one wants to see litigated . # "
say . " # Instead of confrontation , they learned to	intervene	in a loving	, caring way . In the classic grocery situation
religious life of the nation and saw that this power should	intervene	in all sectarian	disputes . In one of this most striking insights
" Turks . " # As the West ponders whether to	intervene	in Bosnia ,	leaders on both sides fear this 5-year-old war at
over their passionate belief that the United States should	intervene	in Bosnia .	# When she was named chief delegate to the
But they do n't seem sufficiently worried about that fact to	intervene	in Bosnia .	It is preposterous to lecture Europeans about their
available for comment . # FERC has shown little inclination to	intervene	in California 's	chaotic energy market , however . Federal

Teaching students to use the interface could help facilitate learning how to choose collocates and prepositions.

COCA Syntax Search Cheat Sheet: Simple search and Tags

Syntax	Meaning	Examples (Click to run in PDF)	Sample matches
One "slot" : Make sure there is no space, or it will be interpreted as two consecutive words			
word	One exact word	mysterious	mysterious
[pos] [pos*]	Part of speech (exact) Part of speech (wildcard) [More information]	[vvg] [v*]	going, using find, does, keeping, started
[lemma]	Lemmas (all forms of a word)	[sing] [tall]	sing, singing, sang tall, taller, tallest
[=word]	Synonyms [More information] [New: synonym chains]	[=strong]	formidible, muscular, fervent
[user:list]	Customized lists [More information]	[mark_davies@byu.edu:clothes]	tie, shirt, blouse
word|word	Any of these words	stunning|gorgeous|charming	stunning, charming, gorgeous
xx x?xx x?xx	Wildcard: * = any # letters Wildcard: ? = one letter	un*ly s?ng s?ng*	unlikely, unusually sing, sang, song song, singer, songbirds
-word	NOT (followed by PoS, lemma, word, etc. Most useful for "multiple slot" queries; see below)	-[nn*]	the, in, is
BASIC TAGS	**[nn*]** noun **[v*]** verb **[j*]** adjective **[r*]** adverb [at*] article **[d*]** determiner **[p*]** pronoun **[app*]** possessive **[i*]** preposition **[c*]** conjuction		

	utter.[j*] – utter as an adjective (e.g. utter despair)	utter.[v*] utter as a verb (e.g. utter words)

COCA Syntax Search Cheat Sheet: Collocates

Multiple "slots" : Create sequences of words, using any of the preceding query types. Note that in each case, there is a space between the word "slots" in the query. These are just a few examples, from an unlimited number of combinations. Note on advanced queries involving variable length between words.	
nooks and crannies	nooks and crannies
fast|quick|rapid [nn*]	fast food rapid transit
pretty -[nn*]	pretty smart pretty as (but **not** pretty girl, pretty picture, etc)
[get] her to [v*]	get her to stay got her to sleep
.|,|; nevertheless [p*] [v*] (Notice that punctuation can be used like any "word"; just make sure that it is separated from words by a space)	. Nevertheless it is ; nevertheless he said
[break] the [nn*]	break the law broke the story
[beat].[v*] * [nn*]	beat the Yankees beaten to death
[=gorgeous] [nn*]	beautiful woman attractive wife
[put] on [ap*] [mark_davies@byu.edu:clothes].[n*]	put on her hat putting on my pants

Source: Corpus of Contemporary American English. (n.d) Brigham Young University. Available at http://corpus.byu.edu/

COCA Syntax Search Cheat Sheet: Advanced List of [POS] Parts of Speech Tags

APPGE	possessive pronoun, pre-nominal (e.g. my, your, our)	EX	existential there
AT	article (e.g. the, no)	FO	formula
AT1	singular article (e.g. a, an, every)	FU	unclassified word
		FW	foreign word
		GE	germanic genitive marker - (' or's)

BCL	before-clause marker (e.g. in order (that),in order (to))	IF	for (as preposition)
CC	coordinating conjunction (e.g. and, or)	II	general preposition
CCB	adversative coordinating conjunction (but)	IO	of (as preposition)
CS	subordinating conjunction (e.g. if, because, unless, so, for)	IW	with, without (as prepositions)
CSA	as (as conjunction)	JJ	general adjective
CSN	than (as conjunction)	JJR	general comparative adjective (e.g. older, better, stronger)
CST	that (as conjunction)	JJT	general superlative adjective (e.g. oldest, best, strongest)
CSW	whether (as conjunction)	JK	catenative adjective (able in be able to, willing in be willing to

DA	after-determiner or post-determiner capable of pronominal function (e.g. such, former, same)	MC	cardinal number,neutral for number (two, three..)
DA1	singular after-determiner (e.g. little, much)	MC1	singular cardinal number (one)
DA2	plural after-determiner (e.g. few, several, many)	MC2	plural cardinal number (e.g. sixes, sevens)
DAR	comparative after-determiner (e.g. more, less, fewer)	MCGE	genitive cardinal number, neutral for number (two's, 100's)
DAT	superlative after-determiner (e.g. most, least, fewest)	MCMC	hyphenated number (40-50, 1770-1827)
DB	before determiner or pre-determiner capable of pronominal function (all, half)	MD	ordinal number (e.g. first, second, next, last)
DB2	plural before-determiner (both)	MF	fraction, neutral for number (e.g. quarters, two-thirds)
DD	determiner (capable of pronominal function) (e.g any, some)		
DD1	singular determiner (e.g. this, that, another)		
DD2	plural determiner (these,those)		
DDQ	wh-determiner (which, what)		
DDQGE	wh-determiner, genitive (whose)		
DDQV	wh-ever determiner, (whichever, whatever)		

ND1	singular noun of direction (e.g. north, southeast)	PPX2	plural reflexive personal pronoun (e.g. yourselves, themselves)
NN	common noun, neutral for number (e.g. sheep, cod, headquarters)	PPY	2nd person personal pronoun (you)
NN1	singular common noun (e.g. book, girl)	RA	adverb, after nominal head (e.g. else, galore)
NN2	plural common noun (e.g. books, girls)	REX	adverb introducing appositional constructions (namely, e.g.)
NNA	following noun of title (e.g. M.A.)	RG	degree adverb (very, so, too)
NNB	preceding noun of title (e.g. Mr., Prof.)	RGQ	wh- degree adverb (how)
NNL1	singular locative noun (e.g. Island, Street)	RGQV	wh-ever degree adverb (however)
NNL2	plural locative noun (e.g. Islands, Streets)	RGR	comparative degree adverb (more, less)
NNO	numeral noun, neutral for number (e.g. dozen, hundred)	RGT	superlative degree adverb (most, least)
NNO2	numeral noun, plural (e.g. hundreds, thousands)	RL	locative adverb (e.g. alongside, forward)
NNT1	temporal noun, singular (e.g. day, week, year)	RP	prep. adverb, particle (e.g about, in)
NNT2	temporal noun, plural (e.g. days, weeks, years)	RPK	prep. adv., catenative (about in be about to)
NNU	unit of measurement, neutral for number (e.g. in, cc)	RR	general adverb
NNU1	singular unit of measurement (e.g. inch, centimetre)	RRQ	wh- general adverb (where, when, why, how)
NNU2	plural unit of measurement (e.g. ins., feet)	RRQV	wh-ever general adverb (wherever, whenever)
NP	proper noun, neutral for number (e.g. IBM, Andes)	RRR	comparative general adverb (e.g. better, longer)
NP1	singular proper noun (e.g. London, Jane, Frederick)	RRT	superlative general adverb (e.g. best, longest)
NP2	plural proper noun (e.g. Browns, Reagans, Koreas)	RT	quasi-nominal adverb of time (e.g. now, tomorrow)
NPD1	singular weekday noun (e.g. Sunday)		
NPD2	plural weekday noun (e.g. Sundays)		
NPM1	singular month noun (e.g. October)		
NPM2	plural month noun (e.g. Octobers)		

PN	indefinite pronoun, neutral for number (none)	TO	infinitive marker (to)
PN1	indefinite pronoun, singular (e.g. anyone, everything, nobody, one)	UH	interjection (e.g. oh, yes, um)
		VB0	be, base form (finite i.e. imperative, subjunctive)
		VBDR	were
PNQO	objective wh-pronoun (whom)	VBDZ	was
PNQS	subjective wh-pronoun (who)	VBG	being
PNQV	wh-ever pronoun (whoever)	VBI	be, infinitive (To be or not... It will be ..)
PNX1	reflexive indefinite pronoun (oneself)	VBM	am
PPGE	nominal possessive personal pronoun (e.g. mine, yours)	VBN	been
PPH1	3rd person sing. neuter personal pronoun (it)	VBR	are
PPHO1	3rd person sing. objective personal pronoun (him, her)	VBZ	is
PPHO2	3rd person plural objective personal pronoun (them)	VD0	do, base form (finite)
PPHS1	3rd person sing. subjective personal pronoun (he, she)	VDD	did
PPHS2	3rd person plural subjective personal pronoun (they)	VDG	doing
PPIO1	1st person sing. objective personal pronoun (me)	VDI	do, infinitive (I may do... To do...)
PPIO2	1st person plural objective personal pronoun (us)	VDN	done
PPIS1	1st person sing. subjective personal pronoun (I)	VDZ	does
PPIS2	1st person plural subjective personal pronoun (we)	VH0	have, base form (finite)
PPX1	singular reflexive personal pronoun (e.g. yourself, itself)	VHD	had (past tense)
		VHG	having
		VHI	have, infinitive
		VHN	had (past participle)
		VHZ	has
		VM	modal auxiliary (can, will, would, etc.)
		VMK	modal catenative (ought, used)
		VV0	base form of lexical verb (e.g. give, work)
		VVD	past tense of lexical verb (e.g. gave, worked)
		VVG	-ing participle of lexical verb (e.g. giving, working)
		VVGK	-ing participle catenative (going in be going to)
		VVI	infinitive (e.g. to give... It will work...)
		VVN	past participle of lexical verb (e.g. given, worked)
		VVNK	past participle catenative (e.g. bound in be bound to)
		VVZ	-s form of lexical verb (e.g. gives, works)
		XX	not, n't
		ZZ1	singular letter of the alphabet (e.g. A,b)
		ZZ2	plural letter of the alphabet (e.g. A's, b's)

Source: Corpus of Contemporary American English. (n.d) Brigham Young University. Available at
http://corpus.byu.edu/

Corpus Glossary:

Token – raw frequency of an item

Tags – symbols (letter and signs) that represent parts of speech

Normalized Frequency – the proportion of tokens per million (…or a thousand, or 10 million) words

Collocates - words nearby a given word (e.g. thick *glasses*, thick *hair*, thick *smoke*)

POS – parts of speech **(e.g.** nouns**,** adjectives and such)