

## English Language Benchmarks

The COCA dataset .....	1
Sentence length .....	2
Language concreteness .....	2
Clause distribution .....	2
Noun density & types.....	2
Verb density & types .....	2
Active and passive forms .....	3
Nominalization.....	3
“Junk” words/Stop words .....	3
References .....	3
TIPS .....	4

### The COCA dataset

From the website of *Corpus of Contemporary American English (COCA)*

<https://www.wordfrequency.info/intro.asp> we read: “The Corpus of Contemporary American English (COCA) is the most widely-used corpus in the world. In March 2020 it was updated for the last time (with data up through Dec 2019), and the word frequency data from the corpus was updated in April 2020. The following are the major changes and improvements in the word frequency data. ... This site contains what is probably the most accurate word frequency data for English. The data is based on the one billion word Corpus of Contemporary American English (COCA) -- the only corpus of English that is large, up-to-date, and balanced between many genres.”

Basically, the COCA corpus, provides a frequency distribution of thousands of English-language lemmatized words by their POS tag value (Part of Speech tag), namely, verb, noun, adjective,..., and genre of text.

“The most basic data shows the frequency of each of the top 60,000 words (lemmas) in each of the eight main genres in the corpus. Unlike word frequency data that is just based on web pages, the COCA data lets you see the frequency across genre, to know if the word is more informal (e.g. blogs or TV and movies subtitles) or more formal (e.g. academic). The following are just a few entries of words at different frequency levels (rank), 1-60,000.”

rank	lemma	PoS	freq	# texts	disp	BLOG	WEB	TV/M	SPOK	FIC	MAG	NEWS	ACAD
614	describe	v	159521	81551	0.94	17718	25573	4270	15796	7866	20583	18065	49650
615	guess	v	159454	74761	0.96	21706	15355	57719	28378	23413	5878	5453	1552
616	choice	n	159277	82417	0.98	28879	23742	17114	16776	10416	21726	17835	22789
617	source	n	158588	74656	0.95	23426	28366	5171	11870	4764	26220	18954	39817
618	mom	n	158511	44697	0.95	12884	10313	66877	19934	25394	13766	8432	911
619	soon	r	158194	95115	0.98	18711	19451	26696	14647	29098	22532	17933	9126
620	director	n	158028	79105	0.94	14248	18521	5554	19383	5197	28196	51981	14948
15024	redhead	n	1766	1209	0.90	96	101	432	86	761	167	95	28

## Sentence length

Modern English is characterized by an **average sentence length between 12 and 25 words**, depending upon corpus type and corpus size; see Kučera and Francis (1967:368–405) and Francis and Kučera (1982:549–553); for a critique, at least with regard to word frequency in relation to corpus size, see Brysbaert and New (2009).

## Language concreteness

Brysbaert et al. (2014) provide measures of concreteness of some 40,000 English words. Using these ratings, Hills and Adelman (2015) show a **long-term trend toward concreteness** in English language use.

## Clause distribution

The frequency distribution of clausal tags tells us something about sentence complexity and style. For Ashok et al. (2013: 1758) “more successful books involve more clausal tags that are necessary for **complex sentence structure and inverted sentence structure** (SBAR, SBARQ and SQ) whereas less successful books rely more on simple sentence structure (S).” Jautze et al. (2013) reach similar conclusions whereby sentence complexity is the discriminant feature between literary novels and chick lit. Literary novels make lesser use of diminutives and have a higher number of relative clauses, of prepositional phrases (PPs), especially PP-adjuncts, and of noun phrases (NPs) – all indicative of the descriptive language more typical of literary novels, than chick lit (Jautze et al. 2013: 77, 79). Churchill, who loved long sentences, was a master of the inverted clause (SINV clausal tag) (e.g., “Never was so much owed by so many to so few.”).

## Noun density & types

How many nouns, and what kinds of nouns are used? Singular, plural, proper, improper? **Males** use more nouns than females, particularly concrete nouns and more geographical references (NER LOCATION, COUNTRY, STATE\_OR\_PROVINCE, CITY), references to quantity, illustratives, location adverbs.

See Biber’s claim of ‘involvement-informational’ dichotomy of females’ (involved) and males’ (informational) writing styles (1988: 115).

## Verb density & types

What is the distribution of verbs? And what types of verbs are used? Which verb voice (active/passive), tense (past, present, future, gerundive), modality (as expressed by auxiliary verbs e.g., obligation, permission, possibility)? **Females** use more verbs, particularly auxiliary verbs than males.

See Biber’s claim of ‘involvement-informational’ dichotomy of females’ (involved) and males’ (informational) writing styles (1988: 115).

For a brilliant analysis of the use of gerundives, see Moretti and Pestre (2015).

### *Active and passive forms*

On the distribution of active and passive forms in large corpora of modern English, see Francis and Kučera (1982:554–555).

### *Nominalization*

Nominalization: turning verbs into nouns (e.g., “The *scolding* of the child was inappropriate.”). Biber (1988: 14–19 and Chapter 7) shows passive syntactic sentence structures and nominalization always go hand-in-hand in a variety of text genres.

### *“Junk” words/Stop words*

Fewer than 400 function words account for over half of the words we use in daily speech, considering that the average native English speaker has a vocabulary of over 100,000 words (e.g., Pennebaker et al. 2003: 570; Chung and Pennebaker 2007: 347)

In computer science, function words, variously labeled as junk words or stop words, are typically disregarded from analysis. But pronouns and auxiliaries have been used as markers of gender-based style in the field of stylometry. **Males**, for instance, since they tend to use more nouns, concrete nouns in particular, will have a higher use of determiners (e.g., ‘the’). **Females** use more negations, pronouns – I, you, she, her, their, myself, yourself, herself – prepositions for and with, and conjunction. They use the “I” more frequently than men. **Depressed individuals** also tend to use the “I” more often.

### *References*

- Biber, D. 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.
- Brysbaert, M., New, B. 2009. “Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English.” *Behav. Res. Methods* 41(4), 977–990.
- Brysbaert, M., Warriner, A.B., Kuperman, V. 2014. “Concreteness ratings for 40 thousand generally known English word lemmas.” *Behav. Res. Methods* 46(3), 904–911.
- Chung, Cindy and James Pennebaker. 2007. “The Psychological Functions of Function Words.” In: pp. 343–359, Klaus Fiedler (Ed.), *Social Communication*, New York: Psychology Press.
- Francis, W. N., Kučera H. (with the assistance of Andrew W. Mackie). 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin Company.
- Hills, T.T., Adelman, J.S. 2015. “Recent evolution of learnability in American English from 1800 to 2000.” *Cognition* 143, 87–92.
- Kučera H., & Francis. W.N. 1967. *Computational Analysis of Present-day American English*. Providence, RI: Brown University Press.
- Moretti, Franco and Dominique Pestre. 2015. “BANKSPEAK: The Language of World Bank Reports.” *New Left Review*, Vol. 92, pp. 75–99.

Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G. 2003. "Psychological aspects of natural language use: our words, our selves." *Annu. Rev. Psychol.* 54, 547–577.

### ***TIPS***

TIPS\_NLP\_Style analysis.py

TIPS\_NLP\_Clause analysis.pdf

TIPS\_NLP\_Noun analysis.pdf

TIPS\_NLP\_Verb analysis.pdf

TIPS\_NLP\_Function words analysis.pdf

TIPS\_NLP\_Nominalization.pdf