English Language Benchmarks

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_PROVICE, 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