Style Analysis

Table of contents

Stylometry or Style Analysis: What is it?

"The analysis of authorial style [is] termed stylometry" in computer science (Neal et al. 2017:86:1)1, or in Holmes's earlier definition (1998: 111) "stylometry [is] the statistical analysis of literary style". Stylometry See alsoms to hold the answer to the question "How about style? Can NLP help?" Can stylometry give us the tools to dig deeper into the reviewers' evaluation of authors' styles?

What are some important NLP style measures?

There is a wide range of measures that are used in computational linguistics to assess issues of style.

Hapax legomena, n-grams, Yule's K, shorter words, vowel words

Computational measures of style have ranged from a focus on unique features of a text, such as *hapax legomena* **or once-occurring words**, punctuation, word case (upper or lower), word, sentence, and paragraph length (and indentation), to a focus on more, rather than less, frequent features such as *n-grams* (of both words and individual characters: Kjell et al. 1994; Sarawgi et al. 2011; Stamatatos 2013; Kestemont 2014; Kestemont et al. 2016), *vocabulary richness measures* **or word type/token ratio** (Holmes 1992: 92-94; Zheng et al. 2006 Holmes 1992: 94, *Yule's Characteristic K*), frequency of *"shorter words"* (i.e., two- or three-letter words) and *"vowel words"* (i.e., words beginning with a vowel) (Holmes 1998), and, especially, frequency of *function words*, the set of words represented by pronouns (e.g., I, you, we, mine), articles (e.g., a, the), conjunctions (e.g., and, or), prepositions (e.g., after, in, to, on, with), and auxiliary verbs (e.g., can, would). These words are also known as "junk" words, as opposed to content words represented by nouns, verbs, adjectives.

POS & DEPREL tags, phrasal & clausal tags, sentence complexity

In recent years, computer scientists have expanded the range of stylistic measures. Using the **POS tags** (Part of Speech tags) and **DEPREL** (Dependency Relations) values obtained from the output of modern parsers, they have focused on the frequency distribution of POS tags (Koppel and Schler 2003; Ashok et al. 2013), of the POS tags of function words (Zheng et al. 2006), of POS tags in conjunction with n-grams as a way to capture syntactic properties or function words and POS tags together (Koppel et al. 2002: 404, 408; Mukherjee and Liu 2010), on the frequency of various measures of **sentence complexity** (Pakhomov et al. 2011; Feng et al. 2012; Jautze et al. 2013), including the percentage of sentences **by phrasal and clausal tags** (Ashok et al. 2013; Feng et al. 2012)5, on verb voice (passive and active; Argamon et al. 2003b) and verb tense (e.g., past, present) using the DEPREL values for verbs and nominalization as a way to capture issues of agency.

Different Approaches in Style Analysis

Machine learning techniques, such as **neural network** or Support Vector Machine (SVM) approaches, were the vogue for some time (e.g., Tweedie et al. 1996; Corney et al. 2002; Koppel et al. 2002; Argamon et al. 2003b; Ashok et al. 2013). Sarawgi et al. (2011) use sophisticated machine learning algorithm and a Probabilistic Context free Grammar (PCFG) parser. Using different types of induction algorithms, Feng and Hirst (2014) similarly measure style beyond the level of sentence, contrary to nearly all stylometric work, **looking across sentences** in terms of local coherence.

Algorithmic sophistication does not necessarily mean better results. At least for now, function words and standard character-based or word-based within-sentence N-grams approaches perform

nearly equally well as more sophisticated approaches that come with the price tag of much greater computing resources (for example, having to solve problems of coreference resolution and dependency parsing or having to train machine learning algorithms) (e.g., See also Sarawgi et al.'s candid admission, 2011: 82). In fact, function words and N-grams are the typical performance benchmarks of new algorithms.

Function words & stop words (vs. content words)

Kestemont (2014: 60), in a "position paper … on the use of function words in computational authorship attribution" recommends moving "away from a language's low-frequency features to a language's high-frequency features, which often tend to be function words" or N-grams. Pennebaker, who has championed the use of function words, writes: "In the English language there are fewer than 200 commonly used particles [function words], yet they account for over half of the words we use." (Pennebaker et al. 2003: 570; Chung and Pennebaker 2007: 347) Function words are typically included in lists of "stop words."

No agreed-upon list of stop words

There isn't an agreed-upon list of stop words, but many computational linguistic tools exclude some 300 to 400 stop words from processing and text corpora are routinely pre-processed to remove them. Yet, Pennebaker has shown how function words, rather than content words, are closely associated to a wide range of variables (Chung and Pennebaker 2007: 355; Pennebaker et al. 2003).

Stylometry: Applications of Style Analysis

Stylometry has been highly successful in the three main areas of its application: authorship attribution, verification, and profiling (Neal et al 2017:86:3).

Authorship attribution

Authorship attribution has tackled a number of media-grabbing cases, followed by much controversy, among them, the authorship of *The Book of Mormon* (whether Sidney Rigdon or Joseph Smith is the author) (e.g., Jockers et al. 2008), of *The Cuckoo's Calling* a detective novel published in 2013 (was Robert Galbraith the author or J.K. Rowling under the pseudonym of Robert Galbraith? (e.g., Juola 2015), not to mention the authorship of the Federalist Papers, the 84 papers published anonymously between 1787 and 1788 in support of the Constitution of the United States (who was/were the author(s) among Alexander Hamilton, John Jay, and James Madison? e.g., Kjell et al. 1994; Tweedie et al. 1996).

Males/females: Do men and women use language differently?

Authorship profiling has been a most fruitful area of investigation for stylometry, tackling such questions as: what is the age, gender, or psychological state of the unknown author of a text? Can a computer algorithm produce valid answers? Empirical findings, by whatever algorithmic method, consistently show a differential use of language by men and women. Males and females

differ in their use of pronouns, certain types of noun modifiers. **Females use more verbs (especially auxiliary verbs), while men use more nouns, especially concrete nouns**. Computer scientists use big data to confirm earlier research on gender language carried out by linguists and literary scholars working with corpus databases (e.g., Palander-Collin 1999) and by psychologists working with small ad-hoc sample data typically analyzed with content analysis (e.g., Colley and Todd, 2002, content analysis of emails about travel by 12 male and 15 female undergraduates or Janssen and Murachver 2004, sample of 37 female and 36 male).

Argamon's work

Argamon, working with different groups, has produced impressive evidence on gendered writing style using different types of big data: a large subset of the British National Corpus (BNC) covering a range of different genres (Koppel et al. 2002; Argamon et al. 2003a), a large database of electronic messages of newsgroups postings on different topics (Argamon et al. 2003b), tens of thousands of blogs (Schler et al. 2006; Argamon et al. 2007). The findings are very consistent across methods and data. **Male style** is characterized by a more frequent use of noun specifiers (determiners – a, the, that, these – numbers – one, two, more, some – modifiers, preposition *of*, and pronoun *its*), while **female style** uses more negations, pronouns – I, you, she, her, their, myself, yourself, herself – prepositions *for* and *with*, and conjunction *and*, auxiliary verbs, and quotations, thus introducing other people's voices. As for bloggers, "male and female bloggers blog about different thing and use different blogging styles." (Schler et al. 2006) Male bloggers talk more about politics, technology, and business than female bloggers, who tend to discuss their personal lives, using a more personal writing style, home, religion, romance, fun. Regardless of gender, however, writing style grows increasingly "male" with age.

Pennebaker's work

Pennebaker's work on function words confirms computer scientists' findings on writing style associated with age and gender. "The ways people use function words reflects their linguistic style" (Chung and Pennebaker 2007: 347). Using a data archive of some 100,000 text files of different kinds of writing of some 80,000 people and for nearly 70 million words, Pennebaker shows that **males' writing** contains higher rates of articles and nouns, particularly concrete nouns, while **females** use more verbs (especially auxiliary verbs) and first person singular pronouns (Chung and Pennebaker 2007: 353-354). **Age** also affects style, older people displaying greater cognitive complexity and, "interestingly," using "more future tense and less past tense the older they get" (Chung and Pennebaker 2007: 354). But Pennebaker extends the analysis of function words beyond demographics, arguing that **function words "carry an array of psychological meanings"** (Chung and Pennebaker 2007: 355; See also also Pennebaker and King 1999) and where "pronouns are among the most revealing" since they "may provide insight into people's level of social integration as well as self-focus." (Pennebaker et al. 2003: 570, 569) Depressed people use more "I" references than non-depressed people. "Too much attention to the self [through the use of "I"] is associated with highly negative emotional states such as depression," as shown even by the writing of suicidal poets (Chung and Pennebaker 2007: 352, 351) Yarkoni (2012), on the basis of a small sample of 694 blogs and the LIWC software, also shows the high correlation between personality and language use, choice of words clearly

associated to the Big Five factors of personality: neuroticism, extraversion, openness, agreeableness, and conscientiousness.

NLP and good/bad novels: How do you know when/what is good?

Computer scientists have also tackled Chong's question: "How do we know whether a novel is good or bad?" (Chong 2011: 64) with automatic algorithms that can pin down the linguistic markers of style of successful novels and poetry (Ashok et al. 2013; Jautze et al. 2013; Kao and Jurafsky 2012). Ashok et al. (2013: 1757) apply a variety of NLP techniques – PCFG parser, Phrasal and Clausal tags, POS tags, sentiment and lexical analysis, SVM, Support Vector Machine classifier – to 100 award-winning novels. They conclude: "prepositions, nouns, pronouns, determiners and adjectives are predictive of highly successful books whereas less successful books are characterized by higher percentage of verbs, adverbs, and foreign words."

Sentence complexity: Literary novels vs chick lit

Furthermore, "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)." (Ashok et al. 2013: 1758) Jautze et al. (2013) reach similar conclusions whereby sentence complexity is the discriminant feature between literary novels and chick lit. Zooming into the sentence parse tree, Jautze et al. (2013: 77, 79) show that literary novels vs chick lit have a higher number of relative clauses, of prepositional phrases (PPs), especially PP-adjuncts, and of noun phrases (NPs), which are indicative of the descriptive language more typical of literary novels. Literary novels make lesser use of diminutives.

Language concreteness and imaginative language

Kao and Jurafsky ask a question similar to Chong's: "What makes a poem beautiful?" (2012: 8) To address the question, they "combine computational linguistics with computational aesthetics" and "compare the stylistic and content features employed by award winning poets and amateur poets." (Kao and Jurafsky 2012: 9, 8) They use *PoetryAnalyzer* to study alliteration, assonance, and sound in rhymes along with other more traditional word and sentence based measures and sentiment analysis (e.g., type/token ratio as a measure of vocabulary richness and word frequency as found in Davies's list of top 500,000 most frequent words from the Corpus of Contemporary American English (COCA) (Davies, 2011) as a measure of readability. Jurafsky found a number of stylistic markers that separate out the 100 award-winning poems from the 100 amateur poems they studied: Professional poems have a richer and more concrete vocabulary and fewer rhymes and negative emotional words, but the single "most important indicator of high-quality poetry ... was the frequency of references to concrete objects, their emphasis on "imagery" (Kao and Jurafsky 2012: 8).

Six dimensions of writing style

Coming from a different scholarly tradition, that of corpus linguistics, and closer to Chong's concern with fiction writing, Egbert (2012) analyzed 391 works of fiction for style using Multi-

Dimensional analysis on 78 different linguistic features. The methodology involves obtaining frequency counts for each of the linguistic features under consideration (be it 78 or 67) in each text and applying factor analysis to find the factor loadings (Biber's dimensions) that account for the stylistic variance in the texts. Biber (1988: 115) suggests six dimensions: Informational versus Involved Production, Narrative versus Non-Narrative Concerns, Explicit versus Situation-Dependent Reference, Overt Expression of Persuasion, Abstract Non-Abstract Information, On-line Informational Elaboration.

What is notable here that the linguistic features depend on a combination of POS tag values (e.g., MD for modal verbs) and specific content values for specific POS tags (e.g., can, could, may, might for MD verbs to define the specific type of possibility modal). Egbert found strong evidence that these fictional works cluster in three of Biber's original six dimensions: Thought Presentation versus Description, Abstract Exposition versus Concrete Action, Dialogue versus Narration. Linguistic features stand in negative and positive relations to one another in each of the three dimensions (e.g., Biber, 1988: 14, had shown that nominalization and passive verbs tend to go together in texts).

What Would Rhetoricians Think of This?

Style, in the computational approach to the study of style, is viewed simply as "a set of measurable patterns which may be unique to an author." (Tweedie et al. 1996: 1) "Much of the style of an author is contained in the statistics of N-tuples of letters extracted from a sample of the author's work" (Kjell et al. 1994: 141) "Style is, to a large extent, determined by the most commonly used words," to the set of "fewer than 400 … function words" (Pennebaker et al. 2003: 569; Chung and Pennebaker 2007: 347).

Reducing style to a handful of "junk" words or even to character N-grams is no doubt computationally easy – and, paradoxically, effective in successfully addressing issues of authorship attribution. Yet, it may be overly optimistic if not altogether pretentious. Style, or elocution, writes Vico (1996 [1711-1741]: 107), is "the most important part of this art [rhetoric] to the extent that eloquence has taken its very name from it." Style, in classical rhetoric, is one of the five canons of rhetoric: invention (*inventio*), arrangement (*dispositio*), and style (*elocutio*), together with memory and delivery. "For it is not enough to know what to say, we must also know how to say it," writes Aristotle in *Rhetoric* (III.1). And that is the purview of style, style as the proper choice of words and of proper length, combined and arranged in the proper way in a sentence, to achieve a proper rhythm and sound, and with proper stylistic differences depending upon setting and occasion. Several chapters of Book III are dedicated to style (1 through 12, out of 15 chapters). In Rhetoric, metaphor occupies a central part in Aristotle's conception of style, with two chapters on its pros and cons (III.2, 3). In Poetics, Aristotle goes as far as calling metaphor "the mark of genius" (XXII). "The whole essence of style of a speech is contained in three elements: First, in grammatically correct speech [the realm of grammar, rather than rhetoric]. Second, in figurative expressions. Third, in the amplifications." (Melanchthon, La Fontaine 1968:223 [1542]) Tropes and figures, of which there were more than 200 at the height of the Renaissance to end up with only one, namely metaphor, by the 20th century, are the real nuts and bolts of style (Franzosi 2017; Franzosi and Vicari 2018). One of the best-known rhetorical treatises of the Renaissance, Erasmus's *Copia* (1978 [1512]) dealt in two volumes with

amplification (*copia*) in both *res* (substance) and *verba* (words) and with the rhetorical figures associated with amplification.

Stylometry: Tools provided in the NLP Suite

The NLP Suite offers a wide range of computational measures of styles. The Style Analysis GUI conveniently brings them all together. The options in this GUI open more specialized GUIs where different arguments can be set for each option.

Clausal analysis

You can run clausal analysis from CoNLL_clausal_analysis.py.
See also TIPS_NLP_Clausal analysis.pdf

Function/junk words analysis

You can run function words analysis from CoNLL_function_words_analysis.py.

See also TIPS_NLP_Function Words Analysis.pdf

Noun & verb analysis

You can run noun/verb words analysis from CoNLL_noun_verb_analysis.py.

See also TIPS_NLP_Noun Analysis.pdf
and TIPS_NLP_Verb Analysis.pdf

Visualize sentence structure (via dependency tree)

You can run the visualization tool from CoNLL_clausal_analysis.py to invoke the Java script DependenSee.jar.

See also TIPS_NLP_Sentence complexity.pdf

Sentence complexity

You can run sentence complexity from CoNLL_clausal_analysis.py to invoke the Java script sentence_complexity.jar.

See also TIPS_NLP_Sentence complexity.pdf

Text readability

You can run the Python package textstat (https://pypi.org/project/textstat/) to obtain text readability scores from CoNLL_clausal_analysis.py.

See also TIPS_NLP_Text readability.pdf

N-grams (word & character)

You can run the NLP Suite Python function for n-grams from either statistics_NLP_main.py or NGrams_CoOccurrences_Viewer_main.py.

The NLP Suite also has a Java tool NGrams_CoOccurrences_Viewer.jar that can be invoked from NGrams_CoOccurrences_Viewer_main.py. The Viewer acts in ways very similar to the Google Ngram Viewer (https://books.google.com/ngrams).

See also TIPS_NLP_Ngrams and Word co-occurrences.pdf
and TIPS_NLP_Google Ngram Viewer.pdf

Hapax legomena (once-occurring words)

Hapax legomena are the once occurring words (or collocations, combination of words, such as "coming out"). To obtain a list of these words, simply run the n-grams script to obtain the frequency of single words, bigrams, trigrams…

You can run the NLP Suite Python function for n-grams from either statistics_NLP_main.py or NGrams_CoOccurrences_Viewer_main.py.

See also TIPS_NLP_Ngrams and Word co-occurrences.pdf

Vocabulary richness (word type/token ratio or Yule's K)

You can run the Yule K algorithm for vocabulary richness from file_checker_converter_main.py.

Unusual words (via NLTK)

You can run the NLTK algorithm for unusual words (or spelling checker) (https://www.nltk.org/book/ch02.html) from file_checker_converter_main.py. This will give you a csv list of unusual/misspelled words.

Short words

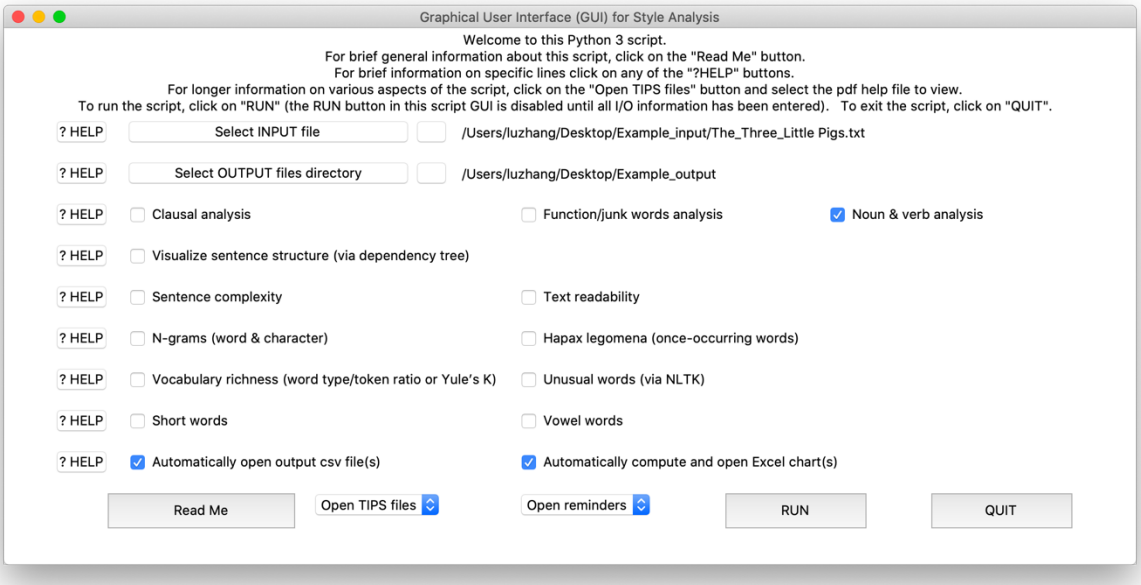
You can run the NLP Suite function for short words from file_checker_converter_main.py. This will give you a csv list of short words (1-4 letters) with their frequency.

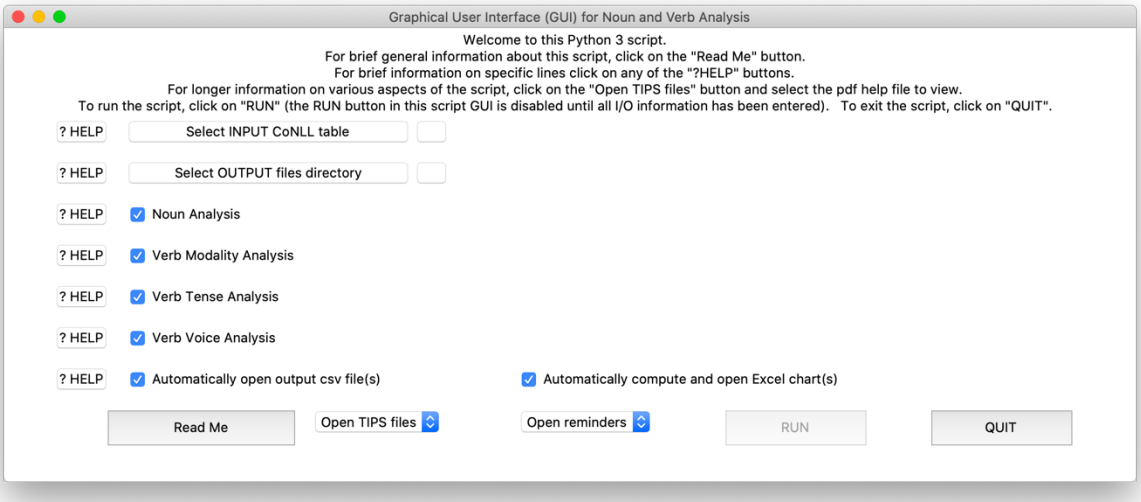
Vowel words (words beginning with a vowel)

You can run the NLP Suite function for vowel words (i.e., words beginning with a vowel) from file_checker_converter_main.py.

This will give you a csv list of vowel words with their frequency.

After selecting the analysis tool(s) you want, click on the RUN button. You will be then redirected to other specialized GUI(s). For instance, if I want to run the Noun & verb analysis, I would select the tool and the new GUI will pop out.





Reference

Abrams, M. H. 1999. *Glossary of Literary Terms.* Seventh edition. New York: Heinle and Heinle. Argamon, Shlomo, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003a. "Gender,
Genre, and Writing Style in Formal Written Texts," *Text*, Vol. 23, No. 3, pp. 321–346.

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2007. “Mining the Blogosphere: Age, Gender and the Varieties of Self-expression.” *First Monday*, Vol. 12, Nos. 9-3, firstmonday.org.

Argamon, Shlomo, Marin Sărić, and Sterling S. Stein. 2003b. “Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results.” In *Proceedings of the 9th ACM SIGKDD*, pp. 475–480. New York: *ACM Press*.

Ashok, Vikas Ganjigunte, Song Feng, and Yejin Choi. 2013. “Success with Style: Using Writing Style to Predict the Success of Novels.” *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1764, Seattle, Washington, USA, 18- 21 October 2013.

Bergsma, Shane, Matt Post, and David Yarowsky. 2012. “Stylometric Analysis of Scientific Articles.” *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 327–337, Montreal, Canada, June 3-8 2012.

Biber, Douglas. 1988. Variation Across Speech and Writing. Cambridge: Cambridge University Press. — 1995. *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Brunner, Annelen. 2013. “Automatic Recognition of Speech, Thought, and Writing Representation in German Narrative Texts,” *Literary and Linguistic Computing*, Vol. 28, No. 4, pp. 563–575.

Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman. 2013. “Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas”. *Behavior Research Methods*, Vol. 46, pp. 904–911.

Burrows, J.F. 1987a. “Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style.” *Literary and Linguistic Computing*, Vol. 2, No. 2, pp. 61-70.

Burrows, J.F. 1987b. *Computation into Criticism: A Study of Jane Austen’s Novels and an Experiment in Method*. Clarendon Press; Oxford University Press.

Childs, Peter and Roger Fowler. 2006 [1973]. *The Routledge Dictionary of Literary Terms.* New York: Routledge.

Chong, Phillipa. 2011. “Reading Difference: How Race and Ethnicity Function as Tools for Critical Appraisal.” *Poetics*, Vol. 39, No. 1, pp. 64-84.

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.

Colley, Anne and Zazie Todd. 2002. “Gender-Linked Differences in the Style and Content of E-Mails to Friends.” *Journal of Language and Social Psychology*, Vol. 21, No. 4, pp.: 380–392. Corney, Malcolm, Olivier de Vel, Alison Anderson, and George Mohay. 2002. “Gender- Preferential Text Mining of E-mail Discourse.” *Proceedings of the 18th Annual Computer Security Applications Conference (ACSAC 2002*).

Cuddon, J. A. Revised by M. A. R. Habib. 2013 [1977]. *A Dictionary of Literary Terms and Literary Theory*. Fifth Edition. Oxford, UK: Wiley-Blackwell.

Davies, Mark. 2011. *Word Frequency Data from the Corpus of Contemporary American English (COCA)*. Downloaded from http://www.wordfrequency.info on May 10, 2011.

Davies, Mark and Dee Gardner. 2010. *Frequency Dictionary of Contemporary American English: Word Sketches, Collocates, and Thematic Lists*. New York: Routledge.
Eder, Maciej, Jan Rybicki, and Mike Kestemont. 2016. "Stylometry with R: A Package for Computational Text Analysis." *The R Journal*, Vol. 8, No. 1.
Egbert, Jesse. 2012. "Style in Nineteenth Century Fiction: A Multi-Dimensional Analysis." *Scientific Study of Literature*, Vol. 2, No. 2, pp. 167–198.
Elson, David K. and Kathleen R. McKeown. 2010. "Automatic Attribution of Quoted Speech in Literary Narrative," In *Proceedings of the Twenty-Fourth AAAI Conference of Artificial Intelligence (AAAI-10)*. Atlanta, GA.
Erasmus, Desiderius. 1978 [1512]. *Copia: Foundations of the Abundant Style (De duplici copia verborum ac rerum commentarii duo)*. *Collected Works of Erasmus*, Vol. 24, Craig R. Thompson (ed.). Toronto: University of Toronto Press.
Escalante, Hugo J., Thamar Solorio, and Manuel Montes-y-Gómez. 2011. "Local Histograms of Character Ngrams for Authorship Attribution." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 288–298.
Feng, Song, Ritwik Banerjee, and Yejin Choi. 2012. "Characterizing Stylistic Elements in Syntactic Structure." In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1522– 1533. Association for Computational Linguistics.
Feng, Vanessa Wei and Graeme Hirst. 2014. "Patterns of Local Discourse Coherence as a Feature for Authorship Attribution." *Literary and Linguistic Computing*, Vol. 29, No. 2, pp. 191-198.
Flood, Barbara J. 1999. "Historical Note: The Start of a Stop List at Biological Abstracts." *Journal of the American Society for Information Science*, Vol. 50, No. 12, p. 1066.
Franzosi, Roberto (ed.). 2017. *Landmark Essays on Tropes and Figures*. New York: Routledge.
Franzosi, Roberto and Stefania Vicari. 2018. "What's in a Text? Answers from Frame Analysis and Rhetoric for Measuring Meaning Systems and Argumentative Structures." *Rhetorica*, Vol. 36, No. 4, pp. 393–429.
Genette, Gerard. 1982. "Rhetoric Restrained", Chapter 6, pp. 103-126, *Figures of Literary Discourse*, Translated by Alan Sheridan, Introduction by Marie-Rose Logan. New York: Columbia University Press.
Herring, Susan C. and John C. Paolillo. 2006. "Gender and Genre Variation in Weblogs." *Journal of Sociolinguistics*. Special Issue: Computer-Mediated Communication, Vol.10, No. 4, pp. 439-459.
Herrmann, J. Berenike, Karina van Dalen-Oskam, and Christof Schöch. 2015. "Revisiting Style, a Key Concept in Literary Studies," *Journal of Literary Theory*, Vol. 9, No. 1, pp. 25–52.
Hills, Thomas T. and James S. Adelman. 2015. "Recent Evolution of Learnability in American English from 1800 to 2000." *Cognition*, Vol. 143, pp. 87–92.
Holmes, David I. 1992. "A Stylometric Analysis of Mormon Scripture and Related Texts." *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, Vol. 155, No. 1, pp. 91-120.
Holmes, David I. 1998. "The Evolution of Stylometry in Humanities Scholarship." *Literary and Linguistic Computing*, Vol. 13, No. 3, pp. 111-117.
Janssen, Anna and Tamar Murachver. 2004. "The Relationship Between Gender and

Topic in Gender-Preferential Language Use." *Written Communication*, Vol. 21, pp. 344–367.

Jautze, Kim, Corina Koolen, Andreas van Cranenburgh, Hayco de Jong. 2013. "From High Heels to Weed Attics: A Syntactic Investigation of Chick Lit and Literature." *Proceedings of the Second Workshop on Computational Linguistics for Literature*, pp. 72–81, Atlanta, Georgia, June 14, 2013.

Jockers, Matthew L., Daniela M. Witten, and Craig S. Criddle. 2008. "Reassessing Authorship of The Book of Mormon Using Delta and Nearest Shrunken Centroid Classification," *Literary and Linguistic Computing*, Vol. 23, No. 4, p. 465–91.

Juola, Patrick. 2015. "The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions." *Digital Scholarship in the Humanities*, Vol. 30, Bo. 1, pp. 100-113.

Kao, Justine and Dan Jurafsky. 2012. "A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry." *Workshop on Computational Linguistics for Literature*, pages 8– 17, Montreal, Canada, June 8, 2012.

Kestemont, Mike. 2014. "Function Words in Authorship Attribution. From Black Magic to Theory?" *Proceedings of the 3rd Workshop on Computational Linguistics for Literature* (CLfL) @ EACL 2014, pp. 59–66, Gothenburg, Sweden, April 27, 2014.

Kestemont, Mike, Justin Stover, Moshe Koppe, Folgert Karsdorp, and Walter Daelemans. 2016. "Authenticating the writings of Julius Caesar." *Expert Systems with Applications*, Vol. 63, pp. 86–96.

Kjell, Bradley, W. Addison Woods, and Ophir Frieder. 1994. "Discrimination of Authorship Using Visualization." *Information Processing and Management*, Vol. 30, No. 1, pp. 141–150.

Klein, Dan and Christopher D. Manning. 2003. "Accurate Unlexicalized Parsing." In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pp. 423–430. Association for Computational Linguistics.

Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni. 2002. "Automatically Categorizing Written Texts by Author Gender." *Literary and Linguistic Computing*, Vol. 17, No. 4, pp. 401–412.

Koppel, Moshe and Jonathan Schler. 2003. "Exploiting Stylistic Idiosyncrasies for Authorship Attribution." In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, Vol. 69, pp. 72-79. CiteSee alsor.

Liddle, Dallas. 2015. "The Coding of Literary Form. Data Mining and the Information Structure of Historical Texts," *IEEE International Conference on Big Data (Big Data)*, pp. 1661-1666, 29 Oct.-1 Nov. 2015, Santa Clara, CA, USA.

Luhn, H.P. 1959. "Keyword in Context Index for Technical Literature (KWIC Index)." Yorktown Heights, NY: IBM, Report RC 127. Also in: 1960. *American Documentation*, Vol. 11, pp. 288–295.

Mamede, Nuno and Pedro Chaleira. 2004. "Character Identification in Children Stories." In pp. pp. 82–90, *EsTAL 2004 Advances in Natural Language Processing, LNCS*. Berlin/Heidelberg: Springer.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, David McClosky. 2014. "The Stanford CoreNLP Natural Language Processing Toolkit." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

Mitton, Roger. 1987. “Spelling Checkers, Spelling Correctors and the Misspellings of Poor Spellers,” *Information Processing & Management.* Vol. 23, No. 5. pp. 495-505.
Moretti, Franco and Dominique Pestre. 2015. “BANKSPEAK: The Language of World Bank Reports.” *New Left Review*, Vol. 92, pp. 75-99.
Mosteller, Frederick and David L. Wallace. 1964. *Applied Bayesian Classical Inference: The Case of the Federalist Papers*. Reading, MA.: Addison-Wesley.
Mukherjee, Arjun and Bing Liu. 2010. “Improving Gender Classification of Blog Authors.” In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP ’10, pp. 207–217, Stroudsburg, PA, USA. Association for Computational Linguistics.
Neal, Tempestt, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. “Surveying Stylometry Techniques and Applications.” *ACM Computing Surveys*, Vol. 50, No. 6, pp. 86:1–86:36, November.
Pakhomov, Serguei, Dustin Chacon, Mark Wicklund, and Jeanette Gundel. 2011. “Computerized assessment of syntactic complexity in Alzheimer’s disease: a case study of Iris Murdoch’s writing”. *Behavior Research Methods*, Vol. 43, No. 1, pp. 136–144.
Palander-Collin, Minna. 1999. “Male and Female Styles in 17th Century Correspondence: I THINK.” *Language Variation and Change*, Vol. 11, No. 2, pp. 123–141.
Pennebaker, James W. and Laura A. King. 1999. “Linguistic Styles Language Use as an Individual Difference,” *Journal of Personality and Social Psychology*, Vol. 77, No.6, pp. 1296-1312. Pennebaker, James W., Matthias R. Mehl, and Kate G. Niederhoffer. 2003. “Psychological Aspects of Natural Language Use: Our Words, Our Selves.” *Annual Review of Psychology*, Vol. 54, pp. 547–77.
Pinker, Steven. 2014. *The Sense of Style: The Thinking Person’s Guide to Writing in the 21st Century*. New York: Penguin Books.
Quinn, Edward. 2006 [1999]. *A Dictionary of Literary and Thematic Terms*. Second edition. New York: Facts on File.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
Ribeiro, Filipe N., Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. “Sentibench-A Benchmark Comparison of State-of-the-Practice Sentiment Analysis Methods.” *EPJ Data Science*, Vol. 5, No. 1, pp. 1-29.
Rudman, Joseph. 1997/1998. “The State of Authorship Attribution Studies: Some Problems and Solutions.” *Computers and the Humanities*, Vol. 31, No. 4, pp. 351-365.
Sarawgi, Ruchita, Kailash Gajulapalli, and Yejin Choi. 2011. “Gender Attribution: Tracing Stylometric Evidence beyond Topic and Genre.” In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pp. 78–86. Association for Computational Linguistics.
Sarmento, Luís and Sérgio Nunes. 2009. “Automatic Extraction of Quotes and Topics from News Feeds.” In *4th Doctoral Symposium on Informatics Engineering*. Pp. 1-12. Porto.
Schler, Jonathan, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. “Effects of Age and Gender on Blogging,” In *Proceedings of the AAAI Spring Symposium Computational Approaches to Analyzing Weblogs*. Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Stanford, California, USA, March 27-29, 2006.

Shutova, Ekaterina. 2010. “Models of Metaphor in NLP.” *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 688–697, Uppsala, Sweden, 11-16 July 2010.

Socher, Richard, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank.” *Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*.

Stamatatos, Efstathios. 2009. “A Survey of Modern Authorship Attribution Methods,” *Journal of the American Society for Information Science and Technology*, Vol. 60, No. 3, pp. 538–556.

Stamatatos, Efstathios. 2013. “On the Robustness of Authorship Attribution Based on Character N-gram Features.” *Journal of Law and Policy*, Vol. 21, No. 2, pp. 421–439.

Sword, Helen. 2012. *Stylish Academic Writing*. Cambridge, Mass: Harvard University Press. Tabata, Tomoji. 1995. “Narrative Style and the Frequencies of Very Common Words: A Corpus- Based Approach to Dickens’s First Person and Third Person Narratives.” *English Corpus Studies*, No. 2, pp. 91-109.

Tweedie, F., S. Singh, and D. Holmes. 1996. “Neural Network Applications in Stylometry: The “Federalist Papers”.” *Computers and the Humanities*, Vol. 30, No. 1, pp. 1-10.

Vico, Giambattista. 1996. *The Art of Rhetoric (Institutiones oratoriae*, 1711–1741), trans. Giorgio A. Pinton and Arthur W. Shippee. Amsterdam: Editions Rodopi B.V.

Yarkoni, Tal. 2012. “Personality in 100,000 Words: A Large-scale Analysis of Personality and Word Use among Bloggers,” *Journal of Research in Personality*, Vol. 44, No. 3, pp. 363–373.

Zheng, Rong, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. “A Framework for Authorship Identification of Online Messages: Writing Style Features and Classification Techniques.” *Journal of the American Society of Information Science and Technology*, Vol. 57, No. 3, pp. 378-393.

TIPS_NLP_Clausal analysis.pdf
TIPS_NLP_Function Words Analysis.pdf
TIPS_NLP_Ngrams and Word co-occurrences.pdf
TIPS_NLP_Noun Analysis.pdf
TIPS_NLP_Sentence complexity.pdf
TIPS_NLP_Text readability.pdf
TIPS_NLP_Verb Analysis.pdf