Zsolt Pál Deli

University of Szeged
Doctoral School in Linguistics
English Applied Linguistics Programme

The lexical analysis of two works by Ernest Hemingway and F. Scott Fitzgerald

https://doi.org/10.48040/PL.2021.16

The scientific field of computational linguistics can significantly contribute to the analysis of literary texts from a variety of perspectives, including educational ones. The purpose of the present study is to investigate and analyze literary texts with the help of computational linguistics devices, with special focus on the difficulty level of vocabulary items, the general vocabulary profile analysis regarding the frequency of occurrence, and the sentence length of selected texts, on which research questions of the present study are based. Ernest Hemingway's work Big Two-Hearted River – Part I., and F. Scott Fitzgerald's The Great Gatsby, were analyzed. It is hypothesized, based on previous research, that the words used in Hemingway's prose will fall into a lower reading difficulty range than that of Fitzgerald's, and that Hemingway's style will generally be simpler than that of Fitzgerald's in terms of syntactic structures and sentence length. Their writings were analyzed with the help of computational linguistics tools. Comparing the text profiles of Hemingway and Fitzgerald for the level of difficulty, it can be concluded that the vocabulary level of their writings is not significantly different. Yet, Fitzgerald's prose contains significantly longer and more elaborate sentences. Language technology appliances may contribute to the critical, detailed and effective analysis of literary works, contributing to other benefits, including language teaching.

Keywords: *computational linguistics, difficulty level, frequency occurrence, sentence length, vocabulary profile*

Introduction

The scientific field of computational linguistics can significantly contribute to the analysis of literary texts from a variety of perspectives, including educational ones. Acquiring quantitative data with the help of computational linguistics devices for the subsequent investigation and analysis of literary texts can lead researchers to a better understanding of literary works; and, this way, this field of applied linguistics can actively promote the progression of literary science as a whole, and literary criticism in particular. Without the help of computational linguistics, the critical analysis of literary works would be an endless task, a practically impossible enterprise to pursue. Therefore, newer and newer developments are created in this field that will potentially

make further advancements to the detailed and effective analyses of literary texts in the future.

The interdisciplinary field of computational linguistics is well elaborated on in Jurafsky's (Jurafsky – Martin, 2008) foundational book on the topic with regard to Natural Language Processing (NLP), text analysis, information retrieval, information extraction, sentiment analysis, corpus linguistics, computational phonology, morphology and syntax, including numerous tagging tools for parts of speech and parsing aids, computational lexicography, machine translation and other NLP applications. The field is developing at a high speed, and quite a large number of researchers are working on both the practical advantages of recent developments and the theoretical aspects of this area of science. Another, similarly accessible book for researchers in the field of electronic analysis of literary texts as well as for linguistic examination of words and phrases in texts is Adolphs' (2006) practical guidance.

For example, electronically stored text archives of corpora that can be used for literary purposes in both quantitative and qualitative statistical text analyses are easy to reach via the internet today. The study of frequency profiles of words and word clusters is not new to the field of computational linguistics. So a number of studies have proven the usefulness of the interpretation of styles of literary texts with the help of the corpus method, or many times "*referred to as corpus stylistics*" (Adolphs, 2006:64).

Characteristic features of authorship can also be described by exploring the frequency profiles of specific words or idiolects by a certain author, attributed to characters in novels, or by using a statistical analysis of frequently occurring words in texts (Burrows, 1987; Hoover, 2002 cited in Adolphs, 2006:64).

It may happen that the assumptions or hypotheses of literary experts are proven or rejected against certain corpora and computational devices, and their discussion may also be influenced by pre-conceived suppositions. Adolphs (2006:65) cites Stubbs (2005) in this respect, referring to Joseph Conrad's Heart of Darkness.

Background

The role of computational linguistics in text analysis

Needless to say, computers serve researchers in many different ways and are an indispensable aid in quantitative research analysis. Linguistics is not an exception to the rule: with the help of computational support, it is possible to conduct rigorous text analysis of literary as well as non-literary pieces, which would have been unimaginable so many years ago. In the old days, surely it

must have been a demanding work on part of the researcher to collect data for quantitative analyses. Then it would have been almost impossible to gain insights into precise quantitative data in a relatively short period of time or deliver findings pointing to certain features of individual artistic styles in literature, and in the process, guessing and intuition were probably part of drawing conclusions as well. However, with the rapid development of information technology, and with the use of computational devices in examining vocabulary profile characteristics, new horizons have been opened for literary inquiry, which may significantly add to scientific research in this field.

Lexical frequency profile

One of the earliest approaches is connected to the frequency examination of learner texts, using Lexical Frequency Profile (Laufer, 1994; Laufer – Nation, 1995; Nation – Waring, 1997), and later by Cobb (1998). These profile instruments were suitable for listing the most frequent words according to categories such as the most common one-thousand words (K1), or the second thousand words (K2), together with presenting the University Word List (UWL), and subsequently the Academic Word List (AWL). Several studies have been conducted, and a number of applications have been put into use, focusing on vocabulary features.

Lexical Frequency Profile (LFP) by Laufer and Nation (1995) is a device designed for measuring lexical richness. The system categorizes vocabulary items according to the various levels for the degree of difficulty. These categories include the first and the second thousand most frequent words (K1 and K2 respectively), including the Academic Word List (AWL), or the University Word List, as it was called earlier (Coxhead, 2000; Coxhead – Nation, 2001 cited in Morris – Cobb, 2004), and words that do not fit into any of the above groupings, the so called off-list words. According to Laufer and Nation (1995), LFP can show a consistent use of vocabulary in several pieces of writings by the same author. One study investigates the sophistication of vocabulary according to the average difficulty level of the individual words by means of their frequency occurrence against a corpus used as a reference (Yoon et al., 2012).

Similarly, the use of vocabulary profile assistance is present in the field of foreign language learning and teaching; therefore, the practice of determining lexical richness in L1 and L2 vocabulary knowledge, both quantitatively and a qualitatively, has been in the center of attention (Kormos – Denes, 2004; Yoon et al., 2012).

Even though we have modern and sophisticated aids at hand at analyzing texts for quantitative parameters, it is still highly unlikely that we can get results with mathematical precision since literature is beyond numbers, and the delicate use of words, phrases and clauses, and the whole network of sentences built into coherent paragraphs is something that is extremely difficult to evaluate merely with numerical devices. This paper only deals with this quantitative approach. Drawing any literary or qualitative conclusions whatsoever is beyond the scope of this minor study; therefore, the focus of the author is of a purely linguistic nature.

Previous studies on Fitzgerald and Hemingway

Rice's study (2016) conducted a quantitative research analysis of Hemingway's prose concerning his style and word choices in comparison to F. Scott Fitzgerald and other contemporaries such as John Steinbeck, Gertrude Stein, and Marcel Proust. The parameters of style, word length, sentence length, lexical richness and the amount of dialogue used in the writings of the mentioned authors were under investigation. The different parts of speech and the occurrence of characteristic words were also discussed, together with an overall evaluation of Hemingway's influence on other writers. The results of the data-driven textual analysis reveal some notable characteristics. For example, in order to prove the assumption that Hemingway's sentences are especially short, the length of his sentences from his works were compared to John Steinbeck's *Grapes of Wrath*, F. Scott Fitzgerald's *The Great Gatsby*, Marcel Proust's *Swann's Way*, and Gertrude Stein's *The Autobiography of Alice B. Toklas*. Hemingway's sentences turned out to be seven words shorter on average, while Proust's sentences proved to be the longest of the authors. One surprising fact was that Steinbeck's novel revealed even shorter sentences than Hemingway's average sentences. Therefore, a more detailed analysis of Hemingway's various pieces of writing was needed in order to have a better understanding of the phenomenon. It is important to note that Hemingway and Steinbeck were contemporaries, but Hemingway's writing career dates back to ten years earlier than Steinbeck's.

The results of Rice's study

The analysis discovered that the early works of Hemingway exhibited shorter sentences than Steinbeck's works, and an increasing progression in the number of words used in his sentences can be detected in Hemingway's prose with time. The later his published works, the longer the sentences. Therefore, Hemingway's younger years present a somewhat different picture than the

works of his later years. Regarding the results for word length, the cluster for the two to six- letter words are quite evenly distributed; however, words consisting of seven letters or more (the ten-dollar words, as Hemingway called them) occur much less in Hemingway's, and even less in Steinbeck's prose, so Steinbeck was, in a way, more Hemingway than Hemingway himself (Rice, 2016). Steinbeck himself acknowledged that a lot of young writers of his age imitated Hemingway, including him. In 1951, William Faulkner said that Hemingway had never used a word that readers were compelled to look up in a dictionary. In A. E. Hotchner's (1966:69-70) *Papa Hemingway: A Personal Memoir*, we can read Hemingway's remark on that: "*He thinks I don't know the ten-dollar words. I know them all right. But there are older and simpler and better words, and those are the ones I use*".

Lexical richness is the proportion of unique words occurring in a given passage, and lower lexical richness indicates more repetition. A given set of unique words can actually be regarded as the author's fingerprint, in a way, his or her own distinct vocabulary. It has been generally concluded that Hemingway's prose indicates a low level, partly manifested in the fact that he uses a repetitive style quite extensively. The study further discusses other aspects, but they are not in the scope of the investigation of the present paper. Yet, the possibility to conduct similar studies from a number of other perspectives with the help of computational linguistics devices may offer researchers an additional potential (Rice, 2016).

Review on Hemingway's and Fitzgerald's style

In 1954, Ernest Miller Hemingway won the Nobel Prize in Literature *"for his mastery of the art of narrative, most recently demonstrated in The Old Man and the Sea, and for the influence that he has exerted on contemporary style"*, reasoned the Nobel committee (NobelPrize.org). Ernest Hemingway and F. Scott Fitzgerald are two of the emblematic figures of the lost generation of artists, who are also referred to as the generation of fire or the WW1 generation. They formed colonies as expatriates in Europe, notably in Paris. They aimed at new forms of expression in stylistic forms, breaking with 19th century, more traditional literary approaches of style, and the appearance of jazz music as a distinctively different art form was also characteristic of that age. The two authors were close acquaintances, very similar in age, and both of them lived in France during the 1920s.

It will not take a long time reading one or more of Hemingway's writings to realize that Hemingway's style is direct, his sentences are generally short, although this approach is sometimes sacrificed for the sake of a balanced diction, using longer sentences as well. He extensively uses

polysyndetons, *"the repeated use of conjunctions to link together a succession of words, clauses or sentences"*, in the form of coordinating structures, especially the conjunction *and* (Baldick, 2004:199) for the special effect of representing a stream of continuity. Hemingway himself makes a remark on the phenomenon in his late work, *The Green Hills of Africa*, when he talks about the style of Dostoevski: *"I wondered if it would make a writer of him, give him the necessary shock to cut the overflow of words and give him a sense of proportion"* (Hemingway, 1963:34).

The diction of short sentences suggests an underlying simplicity of word use, and looking at only the surface level, the false conclusion might be drawn that the words he uses are rather simple. Nevertheless, the vocabulary analysis of his words from the selected text reveals that this assumption is somewhat wrong and is not supported by evidence.

Hemingway's style is direct, simple and lyric with the intellectual element mainly excluded, exploiting the theory of omission, part of the iceberg theory, and it represents a minimalistic writing style concentrating on surface components, and letting the reader to dig under the surface to discover the underlying meaning in an implicit way (the origin of his unique style can be traced back to his background as a cub journalist with the Toronto Star). Apart from his writings being overwhelmingly simplistic, he was also accused of being mannered, and even so simple in his writing *"as an eight-year-old girl"* (Szerb, 1941:866). Even parodies were born in an attempt to imitate his individually unique style, including Fitzgerald himself (Fitzgerald-Turnbull, 1963).

The objective and significance of the study

The purpose of the study is to investigate and analyze literary texts with the help of computational linguistics devices, with special focus on the difficulty level of vocabulary items, the general vocabulary profile, frequency occurrence and sentence length of the selected texts, in the hope that they can point to literary and educational insights, too.

Beyond the literary applications, Cobb (1998) states that frequency lists can be useful in learning new vocabulary. Similarly, Gaetanelle and Granger (2010) confirm the significance of data-driven language learning. Hancioglu and Eldridge (2007) argue for the use of frequency lists for teaching purposes. What is common in the previous applications is the fact that all these are made possible with the help of computational linguistics.

The data used

A passage of Hemingway's *Big Two-Hearted River – Part I.,* first published in *In Our Time*[1]*,* the 1925 edition of a collection of short stories, was analyzed with a selected 3700+ words (Hemingway, 1987:133-143), and a passage from F. Scott Fitzgerald's (1925) *The Great Gatsby,* was analyzed with a selected 3700+ words with the help of computational linguistics devices.

The research questions and hypothesis

Research questions

1. What is the difficulty level of the texts under investigation?
2. What is the length of the sentences under investigation?
3. What is the result of the vocabulary profile analysis of the texts under investigation?

Hypothesis

It is hypothesized, based on previous research, that the words used in Hemingway's prose will fall into a lower reading difficulty range than Fitzgerald's, and Hemingway's style will be simpler than that of Fitzgerald's, in terms of syntactic structures.

Methods

The examination of Hemingway's and Fitzgerald's writings against the ***Oxford Text Checker [tool No. 1.]*** was conducted to see the vocabulary level of their texts. The features of the Oxford Text Checker categorize a typical low intermediate text in such a way that close to 100% of the words form part of the Oxford 3000 keywords. In a typical high intermediate text, 90-95% of the words form part of the Oxford 3000 keywords, and in a typical advanced text, 75-90% of the words form part of the Oxford 3000 keywords.

In a similar way, the examination of some statistical characteristics of the vocabulary profiles of the selected texts was conducted with the help of the ***Compleat Lexical Tutor v. 8.3 [tool No. 2.]***. The Compleat Lexical Tutor v. 8.3 (for data-driven language learning on the web), includes the Vocabprofile, Vocab Stats and Frequency features, among others, which can

[1] "*Give peace in our time, O Lord.*" – a reference to the Evening Prayer in the Book of Common Prayer

be used for statistical analysis to show how many words a certain text contains from the following four frequency levels: (1) the list of the most frequent 1000 word families, (2) the second 1000 word families, (3) the Academic Word List (AWL), and (4) words that do not appear on the other lists.

Results

Hemingway's *Big Two-Hearted River – Part I.,* with 3700+ words analyzed, produced the following results: the words used in the short story fell into the range of 90% in the Oxford 3000 category, which indicates a level of a high intermediate text, while F. Scott Fitzgerald's *The Great Gatsby,* with 3700+ words analyzed, produced the following results: the words used in the short story fell into the range of 87% in the Oxford 3000 category, which indicates a level of an advanced text. Therefore, Hemingway's prose is not significantly different from Fitzgerald's prose according to the word analysis of the Oxford Word Checker.

Comparing the text profiles of Hemingway's and Fitzgerald's writings, with the help of the vocabulary profile, frequency and sentence extractor functions of Compleat Lexical Tutor v. 8.3, the following conclusion can be drawn: regarding the vocabulary level of the two selected writings, and based on the analysis of Lextutor's Vocabprofile and Frequency features, it can be stated that, examining the 3732 tokens for Hemingway's piece of writing, the statistical analysis for all the words used in the text reveals that the types of words he used, indicating frequency, were 878, with the K1 words (0-1000) being 470, and the K2 (1000-2000) being 186 types. The Academic Word List (AWL) words turned out to be 12, and the Off-List Words counted 204 types. Regarding the tokens used in the text, 2967 (79.46 %) fall into the K1, and 380 (10.18 %) into the K2 categories. The figure for the Academic Word List (AWL) is 13 words, which reflects 0.35% of the tokens, and the Oxford Word Checker calculated it as 1%. The number of the Off-List Words is 374 (10.02%).

The same analysis for Fitzgerald's excerpt is the following: examining the 3742 tokens for Fitzgerald's piece of writing, the statistical analysis for all the words used in the text reveals that the types of words he used, indicating frequency, were 1250, with the K1 words (0-1000) being 662, and the K2 (1000-2000) being 177 types. The Academic Word List (AWL) words turned out to be 71, and the Off-List Words counted 318 types. Regarding the tokens used in the text, 3086 (81.60%) fall into the K1, and 219 (5.79%) into the K2 categories. The figure for the Academic Word List (AWL) is 85 words, which reflects 2.25% of the tokens, and the Oxford Word Checker calculated it as 5%. The number of the Off-List Words is 392 (10.36%).

Concerning the sentence length of the two writers, Fitzgerald's prose contains significantly longer sentences as reflected in the figures below. As for Hemingway, the average number of words is 12.23 (SD=9,41), for a total of 304 sentences, while it is 20,39 words for Fitzgerald (SD=15,02), for a total of 183 sentences that contain almost the same amount of words for both writers. Even if we consider the fact that, from a linguistic point of view, the vocabulary level is not significantly different in the writings of the two authors (although it is not true for the academic words), we can draw the conclusion that Hemingway's prose reflects a more direct and simple style, which confirms earlier findings of literary historians, such as Szerb (1941:866). It needs to be noted, nevertheless, that Hemingway consciously created this style himself, and it was his intended purpose to employ such literary practice in his works (Sükösd, 1977; Hemingway, 1963).

Conclusion

From the current study we can draw the conclusion that the scientific field of computational linguistics can significantly contribute to the analysis of literary texts from a variety of perspectives, such as the frequency analysis of words, including measuring sentence length as well as analyzing and checking vocabulary against word lists, together with retrieving and extracting information from texts. Therefore, it can be helpful in a variety of ways, and it may assist us in improving our language skills both in direct and indirect ways. For example, Beatty (2003:7) states that *"a definition of computer assisted language learning (CALL)…is any process in which a learner uses a computer and, as a result, improves his or her language"*. Or CALL can be defined as *"the use of a computer in the teaching or learning of a second or foreign language"* or *"activities which are extensions or adaptations of print-based or classroom based activities…to help achieve educational objectives*" (Richards et al., 1992:73).

Regarding additional language teaching applications, various uses for analyzing texts are suitable to process information on word classes, clause structure and semantic features (Slater et al., 2016). Furthermore, taking advantage of word frequency information can be utilized in the practice of teaching reading, in learner specific syllabus design, calculating with the sequence and average size of vocabulary items, including the use of word families with their derivations and inflections in the course of developing tasks, together with the selection of authentic teaching material considering text difficulty in order for students to successfully cover classroom material according to their level of knowledge (Adolphs, 2006:100).

It should be noted, however, that even if computational linguistics can provide both literary experts and educators with some profitable ideas for further advancements in their fields, the fact that the present study was not carried out on a large scale entails some inherent limitations with it, that is, in order to get more reliable results, larger passages should definitely be included with more parameters to discuss from the literary researcher's perspective, and more practical applications should be considered and exposed from a pedagogical point of view.

References

Adolphs, S. (2006): *Introducing electronic text analysis. A practical guide for language and literary studies*. Routledge: New York. DOI: https://doi.org/10.4324/9780203087701

Baldick, C. (2004): *The Concise Oxford Dictionary of Literary Terms*. Oxford University Press: Oxford

Beatty, K. (2003): *Teaching and Researching Computer Assisted Language Learning*. Longman: New York

Burrows, J. F. (1987): *Computation into Criticism*. Clarendon: Oxford

Cobb, T. (1998): Why & how to use frequency lists to learn words. http://www.lextutor.ca/research/

Coxhead, A. (2000): A new academic word list. *TESOL Quarterly*. 34/2. 213-238. DOI: https://doi.org/10.2307/3587951

Coxhead, A. − Nation, I.S.P. (2001): The specialized vocabulary of English for academic purposes. In: Flowerdew, J. – Peacock, M. (eds.): *Research Perspectives on English for Academic Purposes*. 252-267. Cambridge University Press: Cambridge. DOI: https://doi.org/10.1017/CBO9781139524766.020

Fitzgerald, F. S. − Turnbull, A. (1963): *The Letters of F. Scott Fitzgerald*. Scribner's: New York

Fitzgerald, F. S. (1993): *The Great Gatsby*. Wordsworth Editions Limited: Hertfordshire

Gaetanelle, G. − Granger, S. (2010): How can data-driven learning be used in language teaching? In: O'Keeffe, A. − McCarthy, M. (eds.): *The Routledge Handbook of Corpus Linguistics.* 359–370. Routledge: London. DOI: https://doi.org/10.4324/9780203856949-26

Hancioğlu, N. − Eldridge, J. (2007): Texts and frequency lists: some implications for practising teachers. *ELT Journal.* 61/4. 330-340. DOI: https://doi.org/10.1093/elt/ccm051

Hemingway, E. (1925): *In Our Time*. Boni & Liveright: New York

Hemingway, E. (1963): *Green Hills of Africa*. Scribner's: New York

Hemingway, E. (1987): *The Complete Short Stories of Ernest Hemingway*. Scribner's: New York

Hoover, D. L. (2002): Frequent word sequences and statistical stylistics. *Literary and Linguistic Computing*. 17/2. 157–80. DOI: https://doi.org/10.1093/llc/17.2.157

Hotchner, A. E. (1966): *Papa Hemingway: A Personal Memoir*. Random House: New York

Jurafsky, D. – Martin, J.H. (2008): *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall: New Jersey

Kormos, J. − Denes, M. (2004): Exploring measures and perceptions of fluency in the speech of second language learners. *System*. 32. 145–164. DOI: https://doi.org/10.1016/j.system.2004.01.001
Laufer, B. (1994): The lexical profile of second language writing: Does it change over time? *RELC Journal*. 25/2. 21-33. DOI: https://doi.org/10.1177/003368829402500202
Laufer, B. − Nation, I.S.P. (1995): Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*. 16/3. 307-322. DOI: https://doi.org/10.1093/applin/16.3.307
Morris, L. − Cobb, T. (2004): Vocabulary profiles as predictors of the academic performance of TESL trainees. *System* 32/1. 75-87. DOI: https://doi.org/10.1016/j.system.2003.05.001
Nation, I. S. P. − Waring, R. (1997): Vocabulary size, text coverage and word lists. In: Schmitt, N. − McCarthy, M. (eds.) *Vocabulary: Description, Acquisition, Pedagogy*. 6-19. Cambridge University Press: New York
Rice, Justin. (2016): What Makes Hemingway Hemingway? *LitCharts*.
Richards, J. C. − Platt, J. − Platt, H. (1992): *Longman Dictionary of Language Teaching and Applied Linguistics*. Longman: Harlow, Essex, England
Slater, S. et al. (2016): Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics*. 42/1. 85-106. DOI: https://doi.org/10.3102/1076998616666808
Stubbs, M. (2005): Conrad in the computer: examples of quantitative stylistic methods. *Language and Literature*. 14/1. 5–24. DOI: https://doi.org/10.1177/0963947005048873
Sükösd, M. (1977): *Hemingway világa*. [The World of Hemingway]. Európa Könyvkiadó: Budapest
Szerb, A. (1941): *A világirodalom története*. [The History of World Literature]. Magvető Könyvkiadó: Budapest
The Church of England. (1955): *The Book of Common Prayer and Administration of the Sacraments and Other Rites and Ceremonies of the Church: Together with the Psalter or Psalms of David according to the use of the Anglican Church*. Cambridge University Press: Cambridge
Yoon, S. − Bhat, S. − Zechner, K. (2012): Vocabulary profile as a measure of vocabulary sophistication. *Proceedings of the Seventh Workshop on Innovative Use of NLP for Building Educational Applications*. 180-189. Association for Computational Linguistics: Montreal

Internet sources

How LFP works: http://kojima-vlab.org/lexical_richness/LFP.html
Compleat Lexical Tutor: https://www.lextutor.ca/freq/eng/
LitCharts: https://www.litcharts.com/analitics/LitCharts-Analitics-Hemingway.pdf
Oxford Dictionary: https://www.oxfordlearnersdictionaries.com/oxford_3000_profiler.html
The Nobel Prize in Literature 1954. NobelPrize.org. Nobel Media AB 2020. Wed. 25 Nov 2020. https://www.nobelprize.org/prizes/literature/1954/summary/
http://xroads.virginia.edu/~Hyper/HEMXXINGXxWAY/ch14.html