ABSTRACT

NATURAL LANGUAGE PROCESSING AND SYNTACTIC DIFFERENTIATION:
A CORPUS CASE STUDY

Melissa Wright, M.A.
Department of English
Northern Illinois University, 2017
Gulsat Aygen and Reva Freedman, Thesis Co-Directors

This study analyzed syntactic structures retrieved from Oscar Wilde's *The Picture of Dorian Gray*. Specifically, the constituents and parts of speech within two types of text in the novel – dialogue and descriptive/explanatory – were examined, with the hypotheses that, between the dialogue and the descriptive texts within the narrative, one type would display longer syntactic structures and more embedded clauses, and that specific conjunctions occur more frequently within structures with these clauses. This study utilized natural language processing (NLP) to investigate syntactic length and frequency of parts of speech in the character dialogue and descriptive passages in this narrative. The hypotheses prove to be true, and I prove that Wilde's character dialogue provides simpler and smaller syntactic structures than the descriptive passages. The findings in this study illustrate the importance of context when studying linguistic features – within a conversation, it may be a subconscious expectation that speakers will employ simpler constructions due to working memory (WM) load; however, when reading a descriptive passage within a written work, such limitations may not apply. The results of this study can enable future researchers to investigate linguistic components specific to an individual's written and oral speech patterns that may indicate linguistic-stylistic intricacies, unconscious conversational syntactic principles, and the process of clause embedding.

NORTHERN ILLINOIS UNIVERSITY
DE KALB, ILLINOIS

MAY 2017

NATURAL LANGUAGE PROCESSING AND SYNTACTIC DIFFERENTIATION:

A CORPUS CASE STUDY

BY

MELISSA WRIGHT

A THESIS SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE

MASTER OF ARTS

DEPARTMENT OF ENGLISH

Thesis Co-Directors:
Gulsat Aygen and Reva Freedman

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

ACKNOWLEDGEMENTS

I would like to earnestly thank my advisor, Dr. Gulsat Aygen, for her sage guidance which has helped me grow into a confident linguist. Because of her, I pursued a computational linguistics thesis and am a more well-rounded academic. I will never forget her wise counsel inside and outside of the classroom. I am honored to call her my mentor and friend.

I would also like to express my gratefulness to my advisor, Dr. Reva Freedman, for guiding me through this research process and teaching me the ins and outs of computer programming along the way. I will always appreciate her patience, her reminder to celebrate each pebble, and the advice and guidance that I received from her while beginning to study computational linguistics.

I am also sincerely appreciative for Dr. Betty Birner's support along the way. From providing extremely helpful comments on this thesis, to having discussions about culinary linguistics, I have truly enjoyed learning from her.

My boss, Gail, also deserves to be acknowledged. She provided writing counsel as well as emotional and mental support throughout my graduate school career. She has genuinely made my academic career at Northern Illinois University an enjoyable one.

I am also forever grateful to my best friend, Colleen for teaching me that every mountain has pebbles. You provide an endless supply of laughter and camaraderie. Thank you for always being a source of reassurance, an unwavering ally, and the best friend someone could have.

Last, but certainly not least, I would like to thank my dear husband, Neil, for encouraging me, providing a hand to hold and a shoulder to lean on, and for never letting me doubt myself. He has always encouraged me to pursue my dreams, discussed my linguistic infatuations, supplied infinite love and inspiration, and offered never-ending laughter. His determination to succeed in his own field has set a fantastic example, and I am forever grateful for his presence in my life. Neil, let's do psycholinguistics every day.

DEDICATION

To Neil

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

LIST OF APPENDICES

Appendix Page

PREFACE

This study analyzed syntactic structures retrieved from Oscar Wilde's *The Picture of Dorian Gray* to investigate syntactic environments. Specifically, the constituents and parts of speech within two types of text in the novel – dialogue and descriptive/explanatory – were examined, with the hypotheses that, between the dialogue and the descriptive texts within the narrative, one type would display longer syntactic structures and more embedded clauses, and that specific conjunctions occur more frequently within constructions with these clauses. A series of Python programs were created to clean up the input and make the text readable for the Stanford Parser. The corpus' sentences were subsequently fed into this parser. The information from the parser provided an illustration of how often certain parts of speech occurred, when they occurred, and at what level of the syntactic structure they occurred. Once this was completed, the information was put into a comma separated values (.csv) file – a format very similar to a spreadsheet – and then fed to a final researcher-created Python program to calculate the statistical significance using a two-tailed t-test. The results from this study can be used to forge a path for future research in the computational linguistics field investigating the cognition of speakers and authors; syntacticians and psycholinguists may also find these results useful, as they will also be able to understand the contexts that instigate extended syntactic structures and perhaps examine why these differences should occur.

LITERATURE REVIEW

Noam Chomsky, in 1970, formulated a way to illustrate the manner in which sentences are constructed within cognition by introducing the X-Bar Theory. By devising this theory, Chomsky illuminated the plethora of mental procedures that take place when sentences are formed before anything is even uttered. In other words, Chomsky conceived a theory which seeks to explain the cognitive processes of sentence formation of speakers – what takes place subconsciously when an utterance is generated within the mind.

Chomsky's theory breaks down sentences into phrases (e.g. noun phrases, verb phrases, etc.), features (e.g. phi features), and movements (e.g. verb raising). As Derrick and Archambault (2009) explain:

> Syntax trees provide an aesthetically pleasing way of demonstrating the structure of grammar … By convention, syntax trees should be compact, as symmetrical as possible, and all the lines to children should originate under the centre of a parent. Syntax trees are not limited to sentences, but apply to linguistic units of all ranks including word morphology. Different tree structures can encode the evidence of different meanings for the same output, as in the word 'unlockable' (pp. 1, 2).

In other words, the syntax trees that resulted from Chomsky's X-Bar Theory provide helpful information that demonstrates the possible meaning and the overall structure of the phrases and clauses that make up utterances.

The X-Bar Theory also demonstrates the hierarchy of phrases – some words and/or phrases have dominance over others in a given sentence, and some words and/or phrases have precedence over others in a given sentence. Another illustration that this theory offers, via the X-Bar structure, is the demonstration of embedded clauses – a term which will be used

synonymously with “subordinate clauses” in this paper. For example, a sentence can have a main clause and an embedded clause, as in, *“It is believed* [*that she will write an extraordinary thesis*]*.”* In this example, the bracketed portion of the sentence is the subordinate clause. The word “that” functions as a complementizer which introduces the clause “she will write an extraordinary thesis.” Sentences also have the possibility of having non-finite embedded clauses, as well: *“She seems* [*to be a wonderful student*]*,”* where the bracketed section is the non-finite – containing an unconjugated and un-tensed verb – embedded clause.

Krieghbaum (2014) explains that a syntactic parse tree is able to help illustrate the intricacy of syntactic structures with embedded clauses.

> Subordinate clauses, or SBARs …, are dependent clauses that require more information for the reader to complete the idea. Subordinate clauses usually begin with a subordinating word and include a relative pronoun … Since subordinate clauses are more complex than normal sentences, these are good ways in helping to measure sentence complexity. (p. 7)

Because of the way that syntactic parse trees are able to elucidate these clauses by providing an illustration of the sentences, it is crucial to examine the information provided by these parses closely.

Speaker Identification

Many of the subgenres of linguistics – including syntax, but also extending to phonology, morphology, and others – provide speakers with specific styles of communication. Certain linguistic features and processes reflect intent (e.g. whether a speaker is asking a question or making a statement) and this can determine word order; these features and processes give linguists and computer programmers alike the tools to identify speakers and authors. Forensic linguists, for example, use these subgenres to identify speakers. Oftentimes, as is frequently the

instance for forensic linguistics, this is done using phonological features (e.g. for the purpose of identifying a speaker for a court case). However, using syntactic characteristics (such as embedded clauses and word order), many linguists have made attempts to recognize patterns within speech or writing in order to determine someone's identity.co For example, Rose (2002) explains that a speaker's cognitive objectives may determine both word choice and syntactic structure:

> A change in the cognitive meaning of an utterance will be represented in its linguistic semantic structure, and result in a change in the selection of a word and/or syntax … Differences in cognitive intent are … the main reason for forensic samples having different words and syntax (p. 291).

In other words, the syntactic structure can provide a picture of what the speaker's, or author's, original purpose was. This assists in illustrating how syntactic structure is able to provide a clearer depiction of the speaker or author, which in turn may demonstrate how examining these stylistic intricacies may aid in identifying a communicator or help to understand a speaker's intent. Syntactic characteristics may also serve to aid linguists and computer scientists in understanding the context in which specific structures are implemented.

Nolan (2001) explains that a speaker's language carries a thumbprint, so to speak, put in place by the speaker. "We are frequently able to identify familiar speakers without seeing them … Most people, if they were to be asked whether it is possible to identify speakers from their speech, would readily answer 'yes'" (p. 1). While Nolan is speaking from the point of view of a phonetician, additional linguistic components (e.g. morphology, syntax, etc.), other than phonological, play a role in being able to identify someone from a speech sample because of their interconnected nature. Such identification is essential to understanding human interaction because it not only illustrates how a person individualizes his or her speech, but it also has the

potential to demonstrate within which contexts specific styles of speech are utilized and appropriate.

Cognitive Implications

Additionally, discourse psychologists have been researching the cognitive processes that supply the platform for comprehension within discourse. According to Graesser, Millis, and Zwaan (1997), there are three types of memory: short-term memory (STM), working memory (WM), and long-term memory (LTM). The authors explain, "As a gross approximation, STM holds the most recent clause being comprehended and WM holds about two sentences. Important information is actively recycled in WM" (p. 174). Ergo, it can be concluded that within a conversation, the WM of an interlocuter is limited within the context of a conversation, and thus those partaking in the conversation would not wish to overload the WM by providing too much information in too complex of a manner. Elman (2009) describes theories on how interlocutors cognitively handle lexical ambiguities in real time by describing the limits of the WM.

> The […] hypothesis was that processing occurs in at least two stages […] Two-stage theories are motivated by assumptions regarding limitations in human working memory and processing capacity. These limitations force reliance on a number of syntactic heuristics in order to make a provisional parse of a sentence as it is being processed. (p. 556)

This serves as an explanation as to why, within a narrative, dialogue syntactic structures may be simpler than explanatory passages: if the author is trying to create a realistic conversation, the ability of the WM within such a conversation would intuitively be taken into consideration during the conception of the conversation. Real-time WM necessitates having to limit processing load, and syntactic structures contribute to that load.

Data Mining for Classification

Investigating language in its natural form has the potential to produce a vast array of data; therefore, it is essential to consider how to extract and analyze this data. Data mining (a concept which will be returned to), is a concept which Witten and Frank (2005) demonstrate by explaining that language identification can be used for document classification purposes. One application for data mining, they point out, is:

> *authorship ascription* in which a document's author is uncertain and must be predicted from the text. Here, the stopwords ["words on a fixed, predetermined list of function words" (p. 310)], not the content words, are the giveaway, because their distribution is author dependent but topic independent. (p. 353)

Certain lexical items, in other words, are able to help organize texts into groups based on the authors' word choice preferences, which present themselves within the documents. This presents another example of the possibility of certain data being interpreted in a way that is useful in identifying linguistic characteristics of authors/speakers.

Stylometry can also play an important role in identification. Stylometry is "the study of measurable features of (literary) style, such as sentence length, vocabulary richness and various frequencies (of words, word lengths, word forms, etc.)" and has existed for several centuries (Eder & Rybicki, 2012, para. 1). This analysis process arises from the idea that "there exist such conscious or unconscious elements of personal style that can help detect the true author of an anonymous text" (Eder & Rybicki, 2012, para. 1). Because syntactic structure can be a useful tool in determining characteristics or features of a speaker or author, it can be concluded that stylometry could include the syntax and overall structure of a speech or text. Moreover, it can be assumed that the cognitive processes of a speaker or author exist within a speech sample, and

that every author has a style, either conscious or unconscious, which presents itself within said sample.

Computational Implications

Juola (2006) explains that the increased sophistication of computer software provides a more straightforward way to determine characteristics or intents of an author of a text. “The development of modern computers and large corpora have made it practical to investigate these questions algorithmically via information retrieval techniques” (p. 238). Because computer software provides a manner to more efficiently investigate stylometry, the use of such software to analyze text has provided linguists and programmers a route to more accurate results and a more efficient way of producing those results.

The rise of such technology and the quotidian use of computers have brought about the ability to process natural languages. Natural language processing (NLP) is defined as “the automatic (or semi-automatic) processing of human language” (Copestake, 2002, p. 4). Essentially, NLP bridges the gap in human-to-computer interaction. “The goal of natural language processing is to allow [communication between humans and machines] so that non-programmers can obtain useful information from computing systems” (DeAngelis, 2014, para. 2). The interest in building the bridge between humans and computers arose in 1950 with the creation of the Turing test, whose creator hypothesized that “a computer [can] be considered intelligent if it [can] carry on a conversation with a human being without the human realizing they [are] talking to a machine” (DeAngelis, 2014, para. 1). Krieghbaum (2014) also explains that NLP is helpful for deconstructing sentences to make them more accessible:

> The use of NLP is what gives us our ability to dissect the sentences into detailed usable structures … a normal sentence can be broken down to its most basic features within

> NLP. Though not restricted to language parsing (computer programming languages are broken down similarly), NLP is very helpful in finding […] individual key components in the English language. (p. 6)

Although NLP can be helpful in these capacities, its capabilities are not limited only to artificial intelligence (AI) or deconstruction; today, one of NLP's chief purposes is to help with data extraction. Data is collected by, and in, many different electronic sources, such as emails, and NLP is able to help computer scientists mine such data (known as unstructured data), and allow humans to understand what data is being collected, what information the data provides, and what the numbers extracted from the data are measuring (DeAngelis, 2014).

Computational Linguistics

With the creation and rise in popularity of technology and computational software, the computer science field has presented interesting possibilities for the field of linguistics. Because computers have their own languages (e.g. the Python language, the Java language), and humans have their own languages, the borders have become increasingly blurred between the two disciplines – computational linguistics is now a sought-after subspecialty which intersects the two fields. When considering the field of linguistics itself, there are certain areas of the subject that have been inching ever closer to the computer science boundary for many years. Raskin (1985) illustrates the overlap by saying:

> Everybody who has had some practical experience in NLP knows that at a certain point one has to describe the morphology, syntax, and semantics of a natural language. Not only does linguistics possess most, if not all, of the knowledge one would need in this situation, but much of it is already pre-formated [sic] and pre-formalized for the NLP person. (p. 269)

Raskin continues by saying that the subsections (or levels, as he refers to them) of linguistics (i.e. pragmatics, semantics, syntax, morphology, and phonology) are important components of

knowledge for NLP. Because linguistics generally aims to understand the cognitive processes of creating speech, and computers generally aspire to quickly interpret speech, Raskin mentions that, often, the goals of linguistics and NLP complement each other:

Linguistics Wants:	NLP Needs:
i. To know … about the complex structure mediating the pairings of sounds (spellings) and meanings in natural language	i. To use the shortest and most reliable way from the spellings to the meanings in the text(s) being processed
ii. To structure linguistic meaning and relate it to context	ii. To understand the text and make all the necessary inferences
iii. To distinguish the various levels of linguistic structure, each with its own elements and relations	iii. To use all the linguistic information which is needed for processing the text(s) without any concern for its source
iv. To draw a boundary between linguistic and encyclopedic information to delimit the extent of linguistic competence and, therefore, the limits of the discipline.	iv. To use encyclopedic information on par with linguistic information, if necessary for processing the text(s)
v. To present its findings formally, preferably as a set of rules in an axiomatic theory	v. To implement the available information in a practically accessible and convenient way
	(pp. 276-277)

Essentially, the overall aim of the field of linguistics is to understand the underlying processes of language production, relate the different levels of linguistics to each other, understand linguistic competence (versus prescribed grammatical knowledge), and present linguistic theories; NLP's overarching purpose is to process linguistic information in the most efficient and accurate way, understand a given sample of text, take into account any linguistic information necessary to process this information, and provide the resulting information in a manageable way that is easily

interpretable. The two sub-disciplines feed off each other's strengths and processes, while simultaneously completing each other's goals.

Although NLP and linguistics share certain qualities, in the 1990s, with more of an emphasis placed on empirical methods of research, the NLP field experienced a shift. The stress placed on empirical research, as opposed to the "introspective generalizations that characterized the Chomsky era which held sway in theoretical linguistics" (Liddy, Hovy, Lin, Prager, Radev, Vanderwende, & Weischedel, n.d., p. 1) provoked a change. According to the authors, the NLP field had a shift in focus; it went from examining the possibilities of using language, to the observation of what natural language actually does. The subject of linguistics remains a field where the cognitive processes underlying language production can be explored, while NLP endures as a field that explores existing speech and text as-is for interpretation.

One of the more recent goals in the field of NLP is to enable computers to not just understand phrases and sentences at the word level, but at the broader discourse or pragmatic level. Liddy et al. (n.d.) explain that a "paradigm shift" in NLP needs to be accomplished in order to utilize the higher levels of linguistics (the authors define syntax, morphology, and phonology as the lower levels, while discourse and pragmatics are labeled as the higher levels):

> The desired paradigm shift would require a system's understanding and production of language that goes beyond literal meaning, that is, from just denotative meaning to connotative meaning. For by staying at the denotative level, systems will not be able to accomplish the true human-level language understanding that is accomplished when two individuals interpret the statements of each other in light of what they have learned as to the thoughts, experience, memories, and knowledge of the other. (p. 6)

In other words, the authors believe that once machines accomplish the task of being able to interpret text at a higher level, they will be closer to mimicking what humans are able to do in everyday conversation. Liddy et al. elaborate on this by explaining:

> [accessing multiple large corpora] is one of the most obvious ways to accelerate progress, because accomplishment of evermore human-like NLP requires that what is annotated in texts become more sophisticated and incorporate the richer, more complex, and more implicit aspects of language. (p. 10)

Manning (2015) quoted the director of the Facebook AI Research Lab in Paris, Yann LeCun, as saying, "The next big step for Deep Learning is natural language understanding, which aims to give machines the power to understand not just individual words but entire sentences and paragraphs" (Manning, 2015, p. 701). The goal of those utilizing NLP techniques has now become a matter of providing computers a route to understanding larger blocks of language, rather than word-by-word understanding. Liddy et al. (n.d.) agree with this analysis by saying that NLP is inching ever closer to the human language border in its scope.

> [There is a] realization that NLP, by the blending of statistical and symbolic methods, together with lexical resources such as WordNet, and syntactic and semantic resources such as Prop Bank, plus the availability of large scale corpora on which to test and evaluate approaches, is gaining ground on the goal of realistic comprehension and production of human-like language understanding. (p. 3)

Thus, NLP's increased human-like comprehension is bridging the gap in human-to-computer interaction, which, as previously mentioned, is of utmost importance in today's computational linguistics domain.

Stanford Parser

To aid in comprehending the underlying process of sentence construction, The Stanford Parser was utilized. The parser is a tool that many computational linguists have started using for breaking down sentences in a computational manner. Created by the Stanford Natural Language Processing Group, the parser is able to provide an illustration of how constituents form syntactic structure.

> [The parser] works out the grammatical structure of sentences, for instance, which groups of words go together (as "phrases") and which words are the subject or object of a verb. Probabilistic parsers use knowledge of language gained from hand-parsed sentences to try to produce the *most likely* analysis of new sentences. (Stanford Parser, 2003)

The Stanford Parser is an example of a statistical parser, which uses probabilistic methods, as mentioned above, of decomposing sentences.

When sentences are decomposed by the Stanford Parser, the words are broken down into phrases and parts of speech. Krieghbaum (2014) conducted an NLP experiment in which the Stanford parser was used for syntactic derivation. He explains why this program is so useful:

> [The Stanford Parser] is one of the most accurate and most recognized natural language parser in the linguistics community. The Stanford Parser is a program that parses natural language into grammatical structures of sentences. In other words, it attempts to break down sentences into their basic parts, from the sentence phrasing to the word structure or parts of speech of each word. The Stanford Parser is a Java-based program that is available publically [sic] and has been proven to be relatively reliable in parsing sentences. (Krieghbaum, 2014, p. 9)

Krieghbaum explains that the Stanford Natural Language Processing Group uses Penn Treebank identifiers to name the parts of speech that are the result of a parse. Below are the tags employed by the Penn Treebank:

Table 1

Treebank Tags

1.	CC	Coordinating conjunction	19.	PRP$	Possessive pronoun
2.	CD	Cardinal number	20.	RB	Adverb
3.	DT	Determiner	21.	RBR	Adverb, comparative
4.	EX	Existential there	22.	RBS	Adverb, superlative
5.	FW	Foreign word	23.	RP	Particle
6.	IN	Preposition/subordinating conj.	24.	SYM	Symbol
7.	JJ	Adjective	25.	TO	to
8.	JJR	Adjective, comparative	26.	UH	Interjection
9.	JJS	Adjective, superlative	27.	VB	Verb, base form
10.	LS	List item marker	28.	VBD	Verb, past tense
11.	MD	Modal	29.	VBG	Verb, gerund or present part.
12.	NN	Noun, singular or mass	30.	VBN	Verb, past participle
13.	NNS	Noun, plural	31.	VBP	Verb, non-3rd pers. sing. pres.
14.	NNP	Proper noun, singular	32.	VBZ	Verb, 3rd person singular pres.
15.	NNPS	Proper noun, plural	33.	WDT	Wh-determiner
16.	PDT	Predeterminer	34.	WP	Wh-pronoun
17.	POS	Possessive ending	35.	WP$	Possessive wh-pronoun
18.	PRP	Personal pronoun	36.	WRB	Wh-adverb

(Penn Treebank)

It should be noted that the tags in the table above come from the original Penn Treebank tags; however, a more descriptive set of annotations was eventually published:

> Following the release of the first Penn Treebank CD-ROM, many users indicated that they wanted forms of annotation richer than those provided by the project's first phase, as well as an increase in the consistency of the preliminary corpus. Some also expressed an interest in a less skeletal form of annotation, expanding the essentially context-free analysis of the current treebank to indicate non-contiguous structures and dependencies. (Taylor, Marcus, & Santorini, 2003, p. 8)

Thus, the Penn Treebank II tag set was born, which can aid in the creation of more descriptive parses. For the purposes of this study, the most important change from the original treebank to Penn Treebank II is the addition of the PRN (parenthetical) tag. When parsing quotations, Bies, Ferguson, Katz, and MacIntyre (1995) explain, "If the quotation is discontinuous, the interruptive material is annotated as a parenthetical" (p. 32). The importance of this tag will be addressed further in the methodology section. The above tags break down the information held within individual syntactic structures. Given the hypotheses stated above, the "coordinating conjunction" tag will be especially helpful for illustrating whether hypothesis 2 is proven to be true.

The Stanford Parser breaks down a sentence into dependencies and grammatical structure in a visual manner similar to the X-Bar Theory mentioned above; the difference in the two structures lies in the theoretical and derivational aspects of X-Bar vs. the more concrete, and therefore more programmable, output of the Stanford Parser. A sample parse from the program, using the sentence, "The strongest rain ever recorded in India shut down the financial hub of Mumbai, snapped communication lines, closed airports and forced thousands of people to sleep in their offices or walk home during the night, officials said today" (StanfordParser.n.d.) is seen below:

```
(ROOT
  (S
    (S
      (NP
        (NP (DT The) (JJS strongest) (NN rain))
        (VP
          (ADVP (RB ever))
          (VBN recorded)
          (PP (IN in)
            (NP (NNP India)))))
      (VP
        (VP (VBD shut)
          (PRT (RP down))
          (NP
            (NP (DT the) (JJ financial) (NN hub))
            (PP (IN of)
              (NP (NNP Mumbai)))))
        (, ,)
        (VP (VBD snapped)
          (NP (NN communication) (NNS lines)))
        (, ,)
        (VP (VBD closed)
          (NP (NNS airports)))
        (CC and)
        (VP (VBD forced)
          (NP
            (NP (NNS thousands))
            (PP (IN of)
              (NP (NNS people))))
          (S
            (VP (TO to)
              (VP
                (VP (VB sleep)
                  (PP (IN in)
                    (NP (PRP$ their) (NNS offices))))
                (CC or)
                (VP (VB walk)
                  (NP (NN home))
                  (PP (IN during)
                    (NP (DT the) (NN night)))))))))
    (, ,)
    (NP (NNS officials))
    (VP (VBD said)
      (NP-TMP (NN today)))
    (. .)))
```

Figure 1: Example parse. (Stanford Natural Language Processing Group)

Because the Stanford Parser is so thorough in its breakdown of utterances, the anticipation is that this tool will be able to provide detailed illustrations of the cognition behind sentences, which is ideal for the current study. Because the parser uses probabilistic methods of

parsing sentences by utilizing a large corpus, it allows for higher accuracy when parsing sentences; it is more efficient to use a tool such as this for a large amount of data; and, because of the exhaustive list of parts of speech used by the Stanford Parser, it provides an effective route to analyzing the data of the extended syntactic configurations that are investigated here.

Application to This Study

The previously mentioned research clearly illustrates the importance of the overlap between linguistics and computer science. Because computational methods provide a more efficient route to understanding sentence formation, as was Chomsky's goal in 1970 when he formulated the X-Bar Theory, one can use these methods and the information they provide in order to better understand what components make up a sentence, how these components come together, why the components come together differently, and what scenarios instigate the differences.

The current study investigates *The Picture of Dorian Gray* by Oscar Wilde; the structural syntactic differences in the descriptive pieces of text versus the character dialogue are examined. Syntactic stylometry, a concept mentioned above, will play a role in identifying syntactic variation within the different contexts in the novel. Although Wilde originally composed this piece in 1890, the author has come to be respected in literary enthusiast circles. Recently, Wilde has returned to the spotlight due to an art exhibition. "Inside," the exhibition displayed in Reading Gaol where Wilde was imprisoned for two years, showcases many unique pieces inspired by Wilde's time in the prison. One of Wilde's most famous works, *De Profundis,* has been read weekly by a different performer since the opening of the exhibit. The 50,000 word

letter takes about six hours to read, yet there has not been a shortage of enthusiasm from the volunteer readers (Barker, 2016).

Because of Wilde's re-entry into public consciousness with the art exhibition, and because of his wide range of works during his lifetime, it is worthwhile to examine this author's language and stylistic intricacies. Not only will studying the syntactic differences in his descriptions versus his characters' speech provide a cognitive glimpse into how specific scenarios justify varied syntactic style, but it has the potential to elucidate the author's world-view.

METHODOLOGY

This study is a quantitative evaluation of the frequency of particular syntactic structures and the illustration of those structures within specific contexts. This project breaks down the levels at which constituents and parts of speech occur in the syntax of Wilde's characters' dialogue and his descriptive/explanatory passages. Each sentence from the corpus *The Picture of Dorian Gray* by Oscar Wilde (1890) was parsed. Researcher-generated Python programs were created for data cleanup and analysis. Python was chosen as the designated programming language because of its ability to handle text well. Second, Python can produce output with less lines of code vs. other programming languages that need more lines of code to produce the same results. Therefore, Python is a more powerful and efficient language than the other languages that could have been used for this study. Furthermore, a Python-specific scientific package was used for data analysis, which further highlights the effectiveness of building the pipeline in Python. The final reason Python was chosen was because much of the base code being built upon, for example the code provided by Krieghbaum (2014), was written in Python, and therefore there was no need to handle translation between different programming languages. Below are the steps taken to perform this quantitative analysis.

Input, Methodological Considerations, and Data Cleanup

Input

The original input for this study was retrieved from the Gutenberg Project website[1] in the Unicode format provided on the site. The original file contained 6,067 lines, 318,788 characters, and 57,673 words. Before utilizing Python programs to clean up the file, the following information was manually changed: colons introducing quotes were changed to periods so that the programs mentioned below would recognize the separation between dialogue and non-dialogue when parsing the text, ellipses were replaced by either periods or commas,[2] and paragraphs in French were removed – the version of the parser used would not have been able to parse another language correctly, nor is another language besides English relevant for the purposes of this study. Quoted poetic stanzas were also removed, as they did not produce information relevant to the hypotheses being investigated.

The input file also included specific punctuation and information that required deletion via the researcher created Python programs mentioned above. Moreover, because the file was fed into the parser sentence by sentence, the file needed to be formatted by sentence rather than page width; to achieve this, periods or other terminators (e.g., !, ?, etc.) that occurred in places other than at the end of a sentence needed to be replaced with either a space or a blank. This way, the Python programs that split the document up by sentence would not recognize an abbreviation

[1] www.gutenberg.org

[2] The Python program, which split the input up by sentence, recognized each period in an ellipsis as a sentence terminator and, therefore, added a new line for every period in an ellipsis; because of this, it was necessary to change ellipses to other similar punctuation marks.

such as “Mr.” as marking the end of a sentence. The only abbreviations with inappropriate terminators were: Mr., Mrs., Dr., and Roman numerals.[3]

In addition to removing inappropriate sentence terminators, chapter titles (formatted as “CHAPTER XII,” for example), page numbers (formatted as [12] or […12], depending on whether there was a new chapter that started midway through a page) and front and back matter with copyright and website information had to be removed. These superfluous sections and pieces were removed also using a series of Python programs created by the researcher.

The final program also created two files: one that contained only sentences to feed into the parser and one that included information regarding whether the sentence was dialogue or descriptive text, and the actual sentences (this is called a “standoff markup” file). The standoff markup file was kept until needed after parsing of the sentences, and then the data produced from the parsing were concatenated with the descriptive sentence information in the second file; this allowed for the descriptive information and the parsed sentence to be investigated at one time. This file was needed because the Stanford Parser cannot handle extraneous data irrelevant to the sentences being parsed.

Methodological Considerations

An issue worth noting is that the file fed into the Stanford Parser was not always able to be split by full sentences. For example, a sentence such as, “‘Oh, there is really very little to tell, Harry,’ answered the young painter; ‘and I am afraid you will hardly understand it. Perhaps you will hardly believe it.’” includes both dialogue and non-dialogue; however, the non-dialogue

[3] The corpus formatted Roman numerals as having a period afterward. For example, “XII.”

portion is not a complete sentence. Therefore, the program split the sentence into the sections below:

> Oh, there is really very little to tell, Harry,
> answered the young painter;
> and I am afraid you will hardly understand it.
> Perhaps you will hardly believe it.[4]

This is the input used for the Stanford Parser. In order to remain true to the novel, it was not feasible to place the non-dialogue portion anywhere else nor was it possible to remove it without encountering an ethical issue of whether to concatenate the two portions of dialogue and possibly compromise the integrity of the syntax tree. Therefore, it was decided that the best course of action was to keep the order intact and merely separate text by whether the text was held within quotes.

Although the corpus was split up by type of text, it was also necessary to decide what dictated a sentence's end. It needed to be determined whether a semicolon marked the end of a sentence; whether ellipses – which were replaced – necessitated replacing with periods, commas, colons, or semicolons; whether introductory punctuation, such as colons, leading up to the French verses or Shakespearean excerpts should remain or be replaced with periods, etc. These issues were decided on an individual basis, since there were only a handful of each case. Many decisions were based on test parses done online on the Stanford Parser website[5] and how the parser treated those punctuation marks in context. For example, if the parser treated an ellipsis as a semicolon in the parse, the ellipsis was replaced with a semicolon. It was also decided that semicolons did not indicate the end of a sentence. Since this study examined a written narrative, orthographical cues helped determine this; as there were many occurrences in the corpus of

[4] Wilde, O. (1890). *The Picture of Dorian Gray.*
[5] http://nlp.stanford.edu:8080/parser/

semicolons occurring right before a non-capitalized conjunction, this indicated a continuation of a thought rather than a termination of one. For example, “I see that Basil is in one of his sulky moods; and I can't bear him when he sulks.”[6] Based on the contexts in which they appeared, semicolons were considered an indication of a pause rather than termination of thought.

Data Cleanup

The series of researcher-generated Python programs, the pipeline, aided in cleaning the data before feeding the final file into the parser. Utilizing the pipeline ensured accuracy and speed. Below is a list of steps taken to clean up the data:

1. The corpus was accessed through a third-party website, The Gutenberg Project,[7] and was downloaded in the Unicode format provided by the website.
2. The corpus was then run through a series of researcher-composed Python programs to clean up the data and format the file in preparation for the utilization of the Stanford Parser, which parses sentences and provides a detailed depiction of parts of speech and the levels at which they occur. The individual program descriptions and their purposes are below:
    a. The first program counted characters, words, and lines in the corpus to provide a baseline number. This allowed the researcher to ensure future programs produced the correct modification of the corpus without deleting any extra material.
    b. The second program counted each type of sentence terminator (e.g. !, ?, **;**, **.**, **,**, etc.).

[6] Wilde, O. (1890). *The Picture of Dorian Gray.*
[7] https://www.gutenberg.org/

c. The following program removed the front matter and back matter from the corpus, which was included by Project Gutenberg. This material included copyright information, website information, and information about the text itself.

d. Next, a program was written to remove chapter titles from the corpus.

e. The subsequent program removed periods after abbreviations (i.e. Mr., Mrs., Dr., and Roman numerals)

f. The next program broke up the corpus by sentence to make it easier to feed the corpus through the Stanford Parser. This program also created two files: 1) the standoff markup file (see Appendix A for an example portion of the markup file) that was used for the comma separated values (.csv) file created later, which labeled a sentence as either "N" for "non-dialogue" or "Q" for "quote" and 2) a second file split by either sentence or text type (i.e. quote or non-dialogue).

g. Another program was written after splitting the text in the previous step to indicate to the parser that the end of a line signified the end of a parse. The previous program left lingering semicolons and colons at the end of certain segments in the output file. Because of this, before the creation of this program, instead of stopping at the end of a line, the Stanford Parser would continue a parse until encountering an exclamation point, question mark, or period, which was not a true test of the hypothesis since sections that were not intended to be analyzed as a single unit were being parsed as a singular syntactic structure. This program replaced colons, semicolons, and commas at the end of a line with a period to indicate to the parser that a parse should terminate at the end of a line, splitting up the parses into the appropriate segments.

3. The second file from the program mentioned in the previous step was used as the input for the Stanford Parser. The Stanford Parser created a separate file with the syntax trees illustrating the syntactic structures of the text and parts of speech tags.
4. The following program was created to aid in identifying the frequency of parts of speech and conjunctions, word length, and length of structures from the parser output. The base for the code in this program was provided by Krieghbaum (2014) and was refined to tailor the output from this program and render it relevant to the hypotheses being considered. This program also generated a .csv file for analysis purposes.
5. A final researcher-created Python program was used to take the resulting .csv file and conduct statistical analysis. This program employed SciPy – a statistical package specific to Python that aids with such analysis. The aim of placing the data into this program was to analyze the syntax of the two types of text and count specific parts of speech and the constituents mentioned in the hypotheses. This provided an illustration of the frequency and significance of the results and how they relate to the previously mentioned working hypotheses. This program calculated the following: standard deviations and variances for all three conjunctions for the two types of text, the sum of the three conjunctions in both types of text, standard deviation and variance for tree height, the tree height mean for both types of text, the mean of word counts for both types of text, and the unpaired two-tailed t-test – which produced p-values – for all three conjunctions, word counts, and tree height.

Because of the non-directional nature of the working hypotheses in this study, a two-tailed t-test was utilized to analyze the results. It was unknown before conducting the study whether the dialogue or the explanatory text would hold more embedded clauses; therefore, the two-tailed

test was more appropriate than a one-tailed t-test. This provided an illustration of the frequency and significance of the results and how they relate to the working hypotheses. The t-test Python program mentioned in point 5 conducted the t-test on both types of text for tree height and conjunction occurrences (i.e. “and”, “but”, and “or”), means for word counts, the sum total of tree height values for both types of text, and appropriate variances and standard deviations. Figure 2 illustrates the pipeline of these programs and their resulting outputs (in Figure 2, a rectangle indicates a program, a rounded square indicates an output file, and an oval indicates a third-party program).

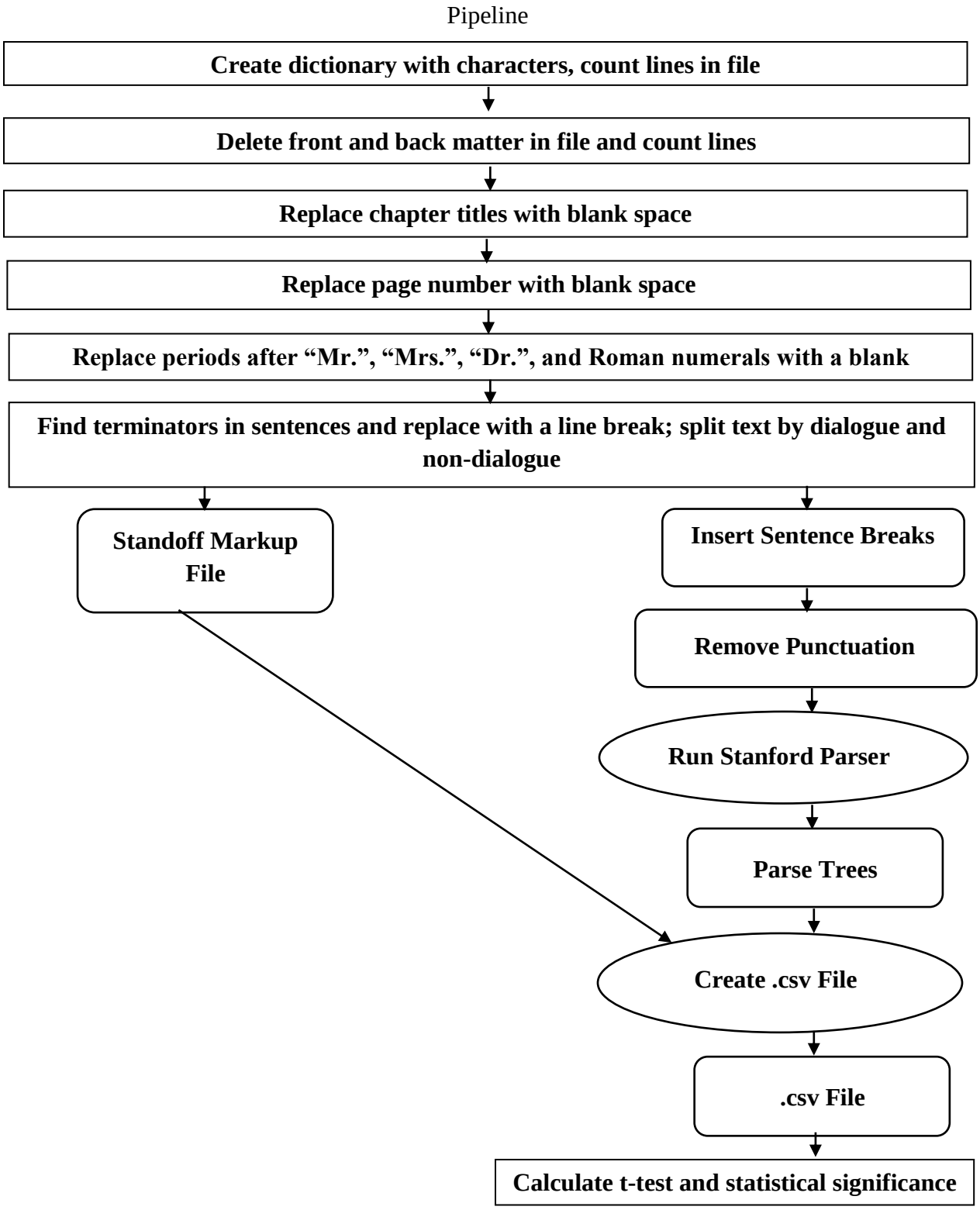

Figure 2: Illustration of Python program data cleanup pipeline**.**

Evaluation

Once the t-test program provided statistical information regarding the parts of speech and tree height, the researcher looked at which outputs were able to provide the most meaningful information based on frequency. As mentioned in the literature section, the updated Penn Treebank II includes tags that were not present in the original version. One of the most important additions, for the sake of this study, is the PRN (parenthetical) tag. When a thought is interrupted by a side thought, the parser applies the PRN tag. Take, for example, "It is a silly habit, I dare say, but somehow it seems to bring a great deal of romance into one's life."[8] The parser takes "I dare say" and attaches a PRN tag to that level of the tree:

```
(ROOT
  (S
    (S
      (NP (PRP It))
      (VP (VBZ is)
        (NP (DT a) (JJ silly) (NN habit))))
    (PRN (, ,)
      (S
        (NP (PRP I))
        (VP (MD dare)
          (VP (VB say))))
      (, ,))
    (CC but)
    (S
      (ADVP (RB somehow))
      (NP (PRP it))
      (VP (VBZ seems)
        (S
          (VP (TO to)
            (VP (VB bring)
              (NP
                (NP (DT a) (JJ great) (NN deal))
                (PP (IN of)
                  (NP (NN romance))))
              (PP (IN into)
                (NP
                  (NP (CD one) (POS 's))
                  (NN life))))))))
    (. .)))
```

Figure 3: Example sentence from *The Picture of Dorian Gray.*

[8] Wilde, O. (1890). *The Picture of Dorian Gray.*

Because of the frequency of this structure in the corpus, it is worth noting the new addition of the PRN tag. Although PRNs were not investigated in this study, because the parse trees illustrate all parts of speech at each level of the syntax, this structure is an interesting outcome.

It is worth noting that the Stanford Parser has limited memory, and two of the sentences in the corpus exceeded that memory. For example, there was a sentence in the N type text that was 198 words long and another, also in the N type text, that was 448 words long (see Appendix B for the 198 word sentence, and Appendix C for the 448 word sentence). These had to be parsed separately due to the memory limitations of the parser.

When comparing the results of the character dialogue vs. the descriptive/explanatory text and the parts of speech, a t-test was performed to determine the statistical significance of the findings. The two-tailed t-test was most appropriate for this research because of the non-directional nature of the hypotheses – it was unknown before the experiment began whether descriptive/explanatory text or dialogue held the longest syntactic structures. The two-tailed t-test provided an illustration of both types of text and the parts of speech and constituents that make up the sentences. Because sentences with subordinate clauses commonly require more words to indicate the embedding of a subsequent clause, and therefore an extended syntax tree, tree height measurements and word counts were used for syntactic analysis to address the hypothesis that between the dialogue and the descriptive texts within the narrative, one type would display longer syntactic structures and more embedded clauses.

This experiment allowed the researcher to analyze the output and determine which of the types of texts held the larger amount of each constituent. The two-tailed t-test was conducted on

each constituent one at a time; because it was assumed that the two types of text being investigated were unequal, the unequal variance component of the t-test was utilized.

RESULTS

To investigate the hypotheses that the N type text produced longer structures with more embedded clauses than the Q type text in *The Picture of Dorian Gray*, the tree height values, word counts, and conjunction values for both types were gathered, and statistical significance testing was done for each category. Because several t-tests were calculated for the two types of text, in order to ensure the results were not skewed, a Bonferroni correction was applied. This made the desired limiting p-value more stringent, because, when multiple t-tests are done for one population, the chances increase of finding a statistically significant finding. In order to safeguard the validity of the statistical significance, the Bonferroni correction was used. In sum, four t-tests were done, and therefore instead of the more common .05 value, the desired p value became .0125. The Stanford Parser created the syntactic parse trees for the sentences in the narrative, which illustrated the intricacies of the trees for both types of text. Figure 4 shows a sample parse from the program:

```
(ROOT
  (S
    (NP (NNP Lord) (NNP Henry))
    (VP
      (VP (VBD stretched)
        (NP (PRP$ his) (JJ long) (NNS legs))
        (PRT (RP out))
        (PP (IN on)
          (NP (DT the) (NN divan))))
      (CC and)
      (VP (VBD shook)
        (PP (IN with)
          (NP (NN laughter)))))
    (. .)))
```

Figure 4: Example parse from *The Picture of Dorian Gray.*

Figure 4 also illustrates the intricacies of the parts of speech and the various levels of the tree structures that are important elements for the purposes of the present study.

The two-tailed t-test was used due to the non-directional nature of the hypotheses. It was unknown before the experiment began whether descriptive/explanatory test or dialogue held the longest syntactic structures with subordinate clauses.

The statistical findings described below are for the segments that the Python programs broke the corpus into to separate the two types of text. Table 2 illustrates the differences in syntax tree height, in word count, for segments within the two types of text analyzed for this study. The t-test determined that there is a difference in tree height between the two types of text and that the N type had the longest syntactic constructions ($p < .000000001$)[9] with a variance of 16.93, which confirms the hypothesis that that the N type text produced more elongated structures than the Q type text, with a variance of 10.89, in *The Picture of Dorian Gray.*

Table 2

Tree Height Information for quotes (Q) and non-quotes (N)

Type	Segment Count	Mean	Standard Deviation
Q	2,967	9.229885	3.300625
N	1,813	10.012700	4.114774

Conjunction counts for both types of text are illustrated in Table 3. In the N type text, 1,269 occurrences of "or" and 1,199 occurrences of "and" were found, compared to 1,069 occurrences of "or" and 558 occurrences of "and" for the Q type text. The least implemented conjunction, "but," occurred only 45 times in the N type text and 104 times in the Q type text.

[9] While the extremely small p-values are surprising for this study, it is assumed that since SciPy was utilized, the correct parameters were applied to ensure accuracy for the values. Future studies may wish to utilize more than one program to calculate the values, however, this is outside the scope of this study.

Table 3

Conjunction Counts for Quotes (Q) and Non-Quotes (N)

Type	Segment Count	AND Sum	BUT Sum	OR Sum
Q	2,967	558	104	1,069
N	1,813	1,199	45	1,269

An interesting finding from this count is that the difference in occurrences of "or" for both types of text is small, however, the difference in occurrences of "and," the second-most-utilized conjunction, for the two types of text is much larger.

This confirms the second hypothesis that specific conjunctions occur more frequently within longer structures with embedded clauses. "Or" was the most utilized conjunction for the N type text, with "and" occurring second-most-frequently. "But" was the least utilized coordinating conjunction in this type but occurred more than twice as much in the Q type text.

Statistical information on the conjunctions within the corpus can be found in Table 4. The three coordinating conjunctions that were looked at for this study are: "and," "but," and "or." The conjunction "or" had the greatest number of occurrences in the N type text, and "but" had the lowest number of occurrences in N type text as well. The N type text had the most occurrences of both "and" ($p < .0000000001$) and "or" ($p < .00000000001$). Table 4 illustrates the statistical information regarding conjunctions in the text.

Table 4

Conjunction Information for Quotes (Q) and Non-Quotes (N)

Type	AND Mean	AND Standard Deviation	BUT Mean	BUT Standard Deviation	OR Mean	OR Standard Deviation
Q	0.206896	0.500111	0.038561	0.194463	0.396366	0.706787
N	0.662065	1.317005	0.024848	0.155662	0.700717	1.193991

Table 5 displays the information for segment lengths, in word count, in the two types of text. The average segment length for the N type text is 15.65 and the average segment length for the Q type text is 9.83 ($p < .000000001$), which serves to further illustrate the difference in syntactic length between the two types of text. The Q type text has a variance of 51.13, and the N type text has a variance of 448.05. This finding illustrates the strength of the confirmation of the first hypothesis that the non-dialogue from *The Picture of Dorian Gray* has lengthened syntactic constructions with more embedded clauses.

Table 5

Segment Length Information for Quotes (Q) and Non-Quotes (N)

Type	Segment Count	Segment Length Mean	Segment Length Standard Deviation	Segment Length Variance
Q	2,967	9.834631	7.151059	51.137651
N	1,813	15.658576	21.167387	448.058278

The differences in overall tree height and occurrences of embedded clauses are not surprising when you take into account the fact that there were two sentences from the N type text

that exceeded the Stanford Parser's memory limitations (see Appendices B and C for the sentences).

DISCUSSION

The original research questions for this study were what environments provoke embedded clauses and whether specific conjunctions occur more frequently in these lengthened structures. When investigating *The Picture of Dorian Gray*, the hypotheses were that, between the dialogue and the descriptive texts within the narrative, one type would display longer syntactic structures and more embedded clauses, and that specific conjunctions would occur more frequently within structures with these clauses. Both hypotheses were proven to be true, with all p-values being significantly less than the $p \leq .0125$ value needed after applying the Bonferroni correction. Therefore, it can be concluded that, within Wilde's narrative, his character dialogue displayed shorter syntactic constructions and his explanatory text employed longer syntactic structures with more embedded clauses, while relying on the conjunction "or" most often within those structures.

Previous literature on NLP has outlined the importance of computational investigative methods for analyzing natural languages, and it has exemplified the potential for increasingly accurate machine comprehension as well as the value of utilizing the linguistics subgenres in such studies. The aim of this study was to exercise these investigative methods within a narrative setting. The results outlined above demonstrate the importance of taking context into account when investigating language – in the case of *The Picture of Dorian Gray*, the character dialogue employed shorter syntactic constructions than the narration. This pattern in the narrative could be due to Wilde's attempt to mirror real-world conversational style: if there are "limitations in

human working memory and processing capacity [which] force reliance on a number of syntactic heuristics in order to make a provisional parse of a sentence as it is being processed" (Elman, 2009, p. 556), then it would be an intuition as an interlocutor to utilize simpler sentences within a discourse. The patterns found within Wilde's work may also display the author's linguistic thumbprint (Nolan, 2001) within his written work. As mentioned, previous research has been done on examining whether a speaker or author can be identified by his or her unique linguistic patterns that are employed within utterances, and this study may indicate that there are indeed such patterns.

Although every sentence in the corpus was examined, there still may remain some limitations. Although much research has been conducted in the past regarding stylometry and cognitive processes in utterance construction, not many studies have delved into syntactic differentiation within a single work by a single author for the purpose of exploring syntactic environmental provocations. Second, only one work was able to be examined. It may be worthwhile in the future to investigate dialogue vs. non-dialogue of more than one text to determine the extent of universality of the findings listed here.

Furthermore, the early steps of the methodology left a bit of room for human judgment. For example, a punctuation mark may not have separated a segment the way the hypothesis required. However, due to the length of the text being examined and the amount of empirical testing done throughout the research process, such occurrences would be minimal and would have very little effect on the results, as the findings were extremely statistically significant.

The findings expressed in this study illustrate the importance of context when studying linguistic features. Within a conversation, it may be a subconscious expectation that speakers will utilize simpler constructions due to working memory (WM) load; however, when reading a

descriptive passage within a written work, such limitations may not apply. The results also display the significance of coordinating conjunctions in relation to the structures with embedded clauses. Because, in this study, “or” was the most utilized conjunction in the N type text – which was found to hold the lengthier syntactic constructions and more embedded clauses – this may highlight the significance of “or” as a coordinating conjunction within longer syntactic structures. The significantly smaller number of “but” conjunctions, occurring only 50 times within the N type segments, may also stress this conjunction’s inability to initiate added information or introduce multiple embedded clauses in one sentence. This may perhaps be due to the conjunction’s pragmatic implication of contrast in meaning between two phrases within a sentence.

In regard to authors’ thumbprints, future research should focus on investigating an author’s works to examine linguistic-structural components and then compare these components to other authors’ works. This will allow linguists to more easily recognize what elements are necessary for something to be able to be deemed a thumbprint and how these thumbprints may transfer into real-world conversations and interactions.

Future research should also conduct discourse analyses to explore syntactic length of utterances within conversation compared to information conveyed that is not expressed within a dialogue to understand length variation between conversational utterances and non-conversational explanatory utterances. Such research has the potential to illustrate the strength behind WM load theories concerning real-time sentence processing, giving discourse psychologists a more solid foundation when investigating cognitive syntactic processing within conversation.

Future investigations could also use discourse analysis to assess syntactic length when interlocutors have different expectations. For example, colleagues discussing research, close friends speaking casually, or acquaintances making small talk all have the potential to produce varied syntactic structures.

Because "but" was by far the least utilized conjunction within either type of text, there is a suggestion that the word's implication of contrast in meaning contributes to its underrepresentation in elongated structures. Future studies should also investigate coordinating conjunctions' role in structures with embedded clauses. Specifically, researchers should investigate whether "but" is universally underrepresented in such constructions, and to what extent its pragmatic "on the contrary" meaning plays a role in this underrepresentation. They should also examine the other coordinating conjunctions' contributions to extended syntactic structures within embedded clauses. Furthermore, because the difference in occurrences of "or" for both the Q type text and the N type text is small and the difference in occurrences of "and," the second-most-utilized conjunction, for the two types is much larger, future research may want to examine whether "and" is universally represented more often in non-quoted material in written texts.

The above analyses and paths to future research have significant implications for the field of linguistics. The results from this study suggest an intuitive behavior to implement simpler syntactic constructions within conversation. They also suggest a reliance on specific coordinating conjunctions within extended structures that contain embedded clauses. Particularly in the context of this study, these sentences seem to rely most heavily on "or," with "and" in a close second, and "but" falling behind in third. This suggests that coordinating conjunctions do play a role in lengthened syntactic configurations with embedded clauses, which necessitates further

investigation into this part of speech. However, perhaps the most significant implication of this study is the possible intuition speakers and authors may have when taking part in or creating a conversation – if interlocutors instinctively know that shorter structures are to be employed when in the context of a conversation, this may indicate underlying unconscious conversational syntactic principles, and this has the potential to further highlight linguistic intuitions and may serve to provide further insight into language processing, linguistic cognition, and unconscious linguistic knowledge.

CONCLUSION

This study investigated Oscar Wilde's syntactic structures in *The Picture of Dorian Gray*, specifically investigating the syntactic variances in the character dialogue and the descriptive/explanatory passages. This research utilized a combination of researcher-composed Python programs, third party programs, and the Stanford Parser to clean up the corpus, analyze the output, and conduct statistical analysis.

The hypotheses that the non-dialogue portions of the narrative would display longer syntactic constructions and more embedded clauses and that specific conjunctions occur more frequently within such constructions are proven to be true.

The results from this study suggest an intuitive behavior to implement simpler syntactic constructions within conversation and a reliance on specific coordinating conjunctions within structures that contain embedded clauses. These implications provide further insight into language processing and unconscious linguistic knowledge, and they indicate the importance of conjunctions when producing syntactic structures with subordinating clauses.

Future researchers wishing to expand upon this topic may wish to conduct discourse analyses to assess syntactic length and embedded clause counts when interlocutors have different conversational goals, investigate coordinating conjunctions' role in lengthier structures with embedded clauses, explore syntactic length of utterances within conversation compared to information conveyed that is not discourse oriented, and examine authors' works to compare

linguistic-structural components between authors and determine what elements are necessary for something to be able to be deemed a linguistic thumbprint.

These possible future research endeavors and the present study present significant implications for both the linguistics and computational linguistics fields: namely, providing further insight into language processing and unconscious linguistic knowledge, identifying conjunctions' role in aiding in the embedding process, and understanding unconscious conversational syntactic principles.

REFERENCES

Alphabetical list of part-of-speech tags used in the Penn Treebank Project. (2003). Retrieved from https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html.

Baker, A. E., & Hengeveld, K. (Eds.). (2012). *Linguistics*. Oxford, UK: Wiley-Blackwell

Barker, V. (2016, October 20). Reading Gaol, where Oscar Wilde was imprisoned, unlocks its gates for art. *NPR*. Retrieved from http://www.npr.org/2016/10/20/498715561/reading-gaol-where-oscar-wilde-was-imprisoned-unlocks-its-gates-for-art

Bies, A., Ferguson, M., Katz, K. & MacIntyre, R. (1995). Bracketing guidelines for Treebank II style: Penn Treebank Project. Retrieved from https://catalog.ldc.upenn.edu/docs/LDC99T42/prsguid1.pdf

Chomsky, Noam. 1970. Remarks on Nominalization. Jacobs, Roderick A. and Rosenbaum, Peter S. (eds.), Readings in English Transformational Grammar, 184-221. Boston: Ginn.

Copestake, Anne. (2002). *Natural Language Processing*. Personal Collection of (Anne Copestake), University of Cambridge, Cambridge, UK. Retrieved from https://www.cl.cam.ac.uk/teaching/2002/NatLangProc/revised.pdf.

DeAngelis, S. F. (2014). The growing importance of Natural Language Processing. *Wired*. Retrieved from http://www.wired.com/insights/2014/02/growing-importance-natural-language-processing/.

Derrick, D., & Archambault, D. (2009). Treeform: Explaining and exploring grammar through syntax trees. *Literary and Linguistic Computing, 25* (1), 53-66. DOI: https://doi.org/10.1093/llc/fqp031

Eder, M., & Rybicki, J. (Eds.). (2012). Proceedings from Digital Humanities 2012: *Introduction to Stylometric Analysis using R.* Hamburg, Germany: University of Hamburg. Retrieved from http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/introduction-to-stylomatic-analysis-using-r.1.html.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science, 33*, pp. 547-582.

Frequently asked questions about Computational Linguistics. (n.d.). Retrieved from http://www.aclweb.org/aclwiki/index.php?title=Frequently_asked_questions_about_Computational_Linguistics#What_is_Computational_Linguistics.3F.

Graesser, A. C., Millis, K. K., Zwaan, R. A. (1997) Discourse comprehension. In Spence, J. T., Darley, J. M., & Foss, D. J. (Eds.). *Annual Review of Psychology, vol. 48,* pp. 163-389.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations, Volume 11, Issue 1*.

Juola, P. (2006). Authorship attribution. In *Foundations and Trends in Information Retrieval*. (Vol. 1, No. 3). Now Publishers.

Krieghbaum, D. (2014). Using machine learning techniques for analyzing educational dialogues and student responses. Northern Illinois University.

Liddy, L., Hovy, E., Lin, J., Prager, J., Radev, D., Vanderwende, L., & Weischedel, R. (n.d.). *Natural Language Processing*. Report one of five for the MINDS workshops: Retrieved from National Institute of Standards and Technology website http://www.itl.nist.gov/iaui/894.02/MINDS/FINAL/NLP.web.pdf.

Lytinen, S. L. (1986). Proceedings from AAAI-86: *Dynamically combining syntax and semantics in Natural Language Processing*. Philadelphia, PA. Retrieved from http://www.aaai.org/Papers/AAAI/1986/AAAI86-097.pdf.

Manning, C. D. (January 7, 2015). The deep learning tsunami. *Computational Linguistics, vol. 41*., No. 4, pp. 701-707. Retrieved from http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00239.

Nolan, F. (2001). *Speaker identification evidence: Its forms, limitations, and roles*. Cambridge, UK.

Raskin, V. (Ed.). (1985). Proceedings from the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages: *Linguistics and Natural Language Processing.* Hamilton, NY: Colgate University. Retrieved from http://www.mt-archive.info/TMI-1985-Raskin.pdf.

Reynolds, D. A. (1995). Automatic speaker recognition using Gaussian mixture speaker models. *The Lincoln Laboratory Journal, 8,* (2), pp. 173-192. Retrieved from https://www.ll.mit.edu/publications/journal/pdf/vol08_no2/8.2.4.speakerrecognition.pdf.

Rose, P. (2002). *Forensic speaker identification*. New York, NY: Taylor & Francis.

Stanford Parser: A statistical parser. (n.d.). Retrieved from http://nlp.stanford.edu/software/lex-parser.shtml.

Taylor, A., Marcus, M., & Santorini, B. (2003). The Penn Treebank: An overview. In A. Abeillé (Ed.), *Treebanks: Building and using parsed corpora* (Vol. 20, pp. 5-22). DOI: 10.1007/978-94-010-0201-1_1

Uszkoreit, H. (2000). *What is Computational Linguistics?* Retrieved from http://www.coli.uni-saarland.de/~hansu/what_is_cl.html.

What is Linguistics? (n.d.). In *SIL International* website. Retrieved from http://www.sil.org/linguistics/what-linguistics.

Witten, I. H., & Frank, E. (2005). Data mining: Practical machine learning tools and techniques. San Francisco, CA: Morgan Kaufmann Publishers.

APPENDIX A

EXCERPT FROM THE STANDOFF MARKUP FILE

Q: Harry,
N: said Basil Hallward, looking him straight in the face,
Q: every portrait that is painted with feeling is a portrait of the artist, not of the sitter.
Q: The sitter is merely the accident, the occasion.
Q: It is not he who is revealed by the painter; it is rather the painter who, on the colored canvas, reveals himself.
Q: The reason I will not exhibit this picture is that I am afraid that I have shown with it the secret of my own soul.
N: Lord Harry laughed.

APPENDIX B

SENTENCE FROM THE N TYPE TEXT WITH 198 WORDS

In a chapter of the book he tells how, crowned with laurel, lest lightning might strike him, he had sat, as Tiberius, in a garden at Capri, reading the shameful books of Elephantis, while dwarfs and peacocks strutted round him and the flute-player mocked the swinger of the censer; and, as Caligula, had caroused with the green-shirted jockeys in their stables, and supped in an ivory manger with a jewel-frontleted horse; and, as Domitian, had wandered through a corridor lined with marble mirrors, looking round with haggard eyes for the reflection of the dagger that was to end his days, and sick with that ennui, that taedium vitae, that comes on those to whom life denies nothing; and had peered through a clear emerald at the red shambles of the Circus, and then, in a litter of pearl and purple drawn by silver-shod mules, been carried through the Street of Pomegranates to a House of Gold, and heard men cry on Nero Caesar as he passed by; and, as Elagabalus, had painted his face with colors, and plied the distaff among the women, and brought the Moon from Carthage, and given her in mystic marriage to the Sun.

APPENDIX C

SENTENCE FROM THE N TYPE TEXT WITH 448 WORDS

Over and over again Dorian used to read this fantastic chapter, and the chapter immediately following, in which the hero describes the curious tapestries that he had had woven for him from Gustave Moreau's designs, and on which were pictured the awful and beautiful forms of those whom Vice and Blood and Weariness had made monstrous or mad: Filippo, Duke of Milan, who slew his wife, and painted her lips with a scarlet poison; Pietro Barbi, the Venetian, known as Paul the Second, who sought in his vanity to assume the title of Formosus, and whose tiara, valued at two hundred thousand florins, was bought at the price of a terrible sin; Gian Maria Visconti, who used hounds to chase living men, and whose murdered body was covered with roses by a harlot who had loved him; the Borgia on his white horse, with Fratricide riding beside him, and his mantle stained with the blood of Perotto; Pietro Riario, the young Cardinal Archbishop of Florence, child and minion of Sixtus IV, whose beauty was equalled only by his debauchery, and who received Leonora of Aragon in a pavilion of white and crimson silk, filled with nymphs and centaurs, and gilded a boy that he might serve her at the feast as Ganymede or Hylas; Ezzelin, whose melancholy could be cured only by the spectacle of death, and who had a passion for red blood, as other men have for red wine,--the son of the Fiend, as was reported, and one who had cheated his father at dice when gambling with him for his own soul; Giambattista Cibo, who in mockery took the name of Innocent, and into whose torpid veins the blood of three lads was infused by a Jewish doctor; Sigismondo Malatesta, the lover of Isotta, and the lord of Rimini, whose effigy was burned at Rome as the enemy of God and man, who strangled Polyssena with a napkin, and gave poison to Ginevra d'Este in a cup of emerald, and in honor of a shameful passion built a pagan church for Christian worship; Charles VI, who had so wildly adored his brother's wife that a leper had warned him of the insanity that was coming on him, and who could only be soothed by Saracen cards painted with the images of Love and Death and Madness; and, in his trimmed jerkin and jewelled cap and acanthus-like curls, Grifonetto Baglioni, who slew Astorre with his bride, and Simonetto with his page, and whose comeliness was such that, as he lay dying in the yellow piazza of Perugia, those who had hated him could not choose but weep, and Atalanta, who had cursed him, blessed him.