LITERATURE, LINGUISTICS & CRITICISM | RESEARCH ARTICLE

A corpus-based stylistic analysis of online suicide notes retrieved from Reddit

Eman Adil Jaafar* and Haya Abdul-Salam Jasim

Abstract: This study examines the stylistic linguistic features of a collection of suicide notes retrieved from the online platform Reddit from the period (2012–2020). It aims to conduct a detailed corpus-based stylistic analysis. Furthermore, it aims at checking specific selected stylistic categories. The quantitative analysis was performed using Wordsmith 8.0 software which helps to identify keywords in context. Keywords are indicators of the style of the language of the selected corpus. The analysis result indicated that the redditors in the online suicide corpus usually included simple words with fairly short sentences. Their writings tend to be less lexically diverse, concentrating on one or two themes with a lower TTR (type-token ratio), which indicated that their words were relatively repetitive. Online suicide notes show high frequency in using first-person singular pronouns, and concerning the distribution of content words, the verbs were the most frequently used part of speech, followed by nouns, adjectives, and adverbs. Furthermore, the keywords list helps to reveal stylistic and linguistic elements.

Subjects: Language & Linguistics; Stylistics; English Language

Keywords: Corpus stylistics; suicide notes; Reddit; Wordsmith tools

1. Introduction

This study is mainly concerned with examining the linguistic stylistic features of the language of suicide notes that have been posted on Reddit. Although there is no particular genre dedicated to suicide, the social expectations of the context (the act of committing suicide) present a particular stereotype related to suicide language. This idea is confirmed by the previous studies when they compare between the writing of genuine and simulated suicidal people (by illustrating the frequent use of first-person singular pronouns and focusing on a particular part of speech than the others).

ABOUT THE AUTHOR

Eman Adil Jaafar is Assistant Professor of linguistics at the Department of English/College of Education for Women/University of Baghdad. Her research interests include but are not restricted to stylistics, corpus and cognitive stylistics and applied linguistics.

Haya Abdul-Salam Jasim a researcher graduated from the Department of English/College of Education for Women/University of Baghdad. Her research interests include Forensic linguistics, Stylistics, and Corpus stylistics.

PUBLIC INTEREST STATEMENT

The paper shows an attempt to analyze the language of suicide notes posted online on one of the social media platforms, Reddit. This is achieved by employing corpus stylistic tools. Qualitative and quantitative methodology is applied to discover stylistic as well as linguistic features of the corpus of suicide notes. The study reveals interesting facts about the online suicide notes.

*Corresponding author: Eman Adil Jaafar, Department of English, University of Baghdad, Baghdad, Iraq E-mail: eman_jafer@coeduw.uobaghdad.edu.iq

Received: 25 December 2021
Accepted: 23 February 2022

© 2022 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.
Osgood and Walker (1959), Edelman and Renshaw (1982), and many others are examples of these types of studies.

Grounded on the statistic presented by the World Health Organization in the United States in 2019 particularly, suicide is regarded as the second leading cause of death to people whose ages range between 25–34, and the third for those whose ages range between 15–25. Suicide can be defined, according to the Collins Concise English Online Dictionary, as “the act or instance of killing oneself intentionally”; meanwhile, a suicidal person is “a person who kills himself or herself intentionally.” As for the “suicide note,” it can be defined as “a note written by someone who intends to kill themselves saying that this is what they are going to do and sometimes explaining why.”

According to Gregory (2018), a suicide note is a message which is written by someone when he decides to end his life to either tell others about his feeling, to express his soreness, or to hold responsible for a particular person. Suicidal people used to write long notes that include more information about their feelings since they will not be able to say it later. A suicidal thought or suicidal ideation is a case that happened to people who have either psychological problems or do not have it at all after suffering from depression. So, a suicidal thought post is a post that is written and posted by someone on social media platforms, like Twitter, Reddit, or any other different platform, after being under stress to ask for help or tell others about their pains and so on. Hence, these platforms are regarded as a window to see the participants’ true feelings or read their last words. Barak and Miron (2005) confirm that suicidal people can be efficiently studied through their online writing. Therefore, suicide notes or suicide thought posts help people understand others’ true feelings from their writing style.

It is assumed that online suicide notes have distinctive language features, for example, the result of Roubidoux’s (2012) study about the use of the pronouns in suicide notes infers a high usage for first-person active pronouns, exclusive pronouns, and singular pronouns, whereas little usage is for passive, inclusive, and plural pronouns. Thus, this current study aims to identify such linguistic features by performing a corpus-based stylistic analysis of a collection of online suicide notes retrieved from Reddit to examine other distinctive stylistic features.

This study followed McIntyre and Walker’s (2019) main ideas in their book Corpus Stylistics theory and practice. to conduct its analysis, it adapts the definition they offer to Corpus Stylistics “the application of theories, models, frameworks from stylistics in corpus analysis”(pp.14–15). It uses methods and tools from corpus linguistics and incorporates them with models and theories from stylistics; this cooperation presents a means to describe, measure, analyse the style of both literary and non-literary texts. To achieve the aims of this study, the data is organised in the form of a corpus consisting of a collection of suicide notes retrieved from Reddit (Suicide Notes Corpus). Then, following Leech and Short’s (2007) method, the researchers stylistically analyse this corpus. To the researchers’ knowledge, there is a gap in the literature in conducting a corpus-based stylistic approach to analyses suicide notes posted on Reddit. However, few studies have tackled the language of suicide notes from other linguistic perspectives. Some of these studies that tackled suicide language are illustrated by Jasim and Jaafar (2022).

2. Method
This study is influenced by the ideas presented by McIntyre and Walker (2019) to conduct a detailed corpus-based analysis of a collection of suicide notes that have been retrieved from Reddit. This study employs both the quantitative and qualitative methods to examine the linguistic features of the data, selected stylistics categories have been chosen from the checklist presented by Leech and Short (2007) to perform the stylistic analysis of this study. This checklist includes four major categories, as illustrated in Figure 1; this study uses selected sub-categories from the lexical and grammatical categories.
3. Procedures
To achieve the objectives of the study, the researchers have performed the following steps:

(1) Examining the sentences and words length in each corpus.
(2) Calculating TTR and STTR; the lexical distribution.
(3) Identifying the most frequently used part of speech in each corpus.
(4) Creating a wordlist to specific word’s classes in each corpus; the content words (nouns, verbs, adjectives, adverbs), and the pronouns used in each corpus.
(5) Analysing these wordlists using selected categories from Leech and Short’s (2007) checklist.
(6) Performing a keywords analysis.

3.1. Corpus tools
In order to conduct the quantitative analysis, two software are involved in this study; the first is WordSmith 8.0 (Scott, 2021), which has a remarkable statistic that makes it different from other software. The second one is TagAnt 1.2 (Anthony, 2015) software which is used to annotate the corpus.

3.2. Data
The data for the current study have been retrieved from Reddit. Unlike the other online websites, Reddit allows its participants to register with fake names, and they do not have to present an email or any other personal information; if the participants add any, these pieces of information remain private. Redditors can vote, post, and comment on public posts or other people’s posts. This platform includes massive numbers of Subreddit, and each one is related to a particular topic; the current study used the post that people tend to post in subreddits (r/Suicidal_Thoughts), (r/Depression), and (SuicideWatch).

The first step is to determine, precisely, what is counted as a suicide note; for this study, a suicide note is any piece of text posted online by a suicidal person on suicide dedicated websites under the
title of “My suicide note,” “My suicide letter,” “A Suicide notes,” and the like. The data have been collected using “Reddit API”, by using python, the system has been instructed to collect the posts from particular subreddits, r/SuicideWatch and r/depression, based on specific keywords in the titles. It should be mentioned that the data have been stored on an internal storage to ensure private access to the data by the researchers only; this is to comply with Reddit terms about the participants’ privacy. Then, the researchers have cleaned the selected data, and this process includes spelling out the abbreviations found in the data and changing the spelling of any UK words into US words by using Microsoft word editor and Grammarly for identifying the words in question.

To sum up, fifty suicide notes (See, Table 1.) have been collected by the researchers and stored in the form of a word-document file with full information about each suicide note, including the account’s names, the subreddit in which the note is posted, the date, and the retrieval date. All the accounts have been checked continuously for four months (for any new activity status). Thus, all the accounts are proven to be offline after posting their final post announcing their willingness and decisions to commit suicide.

In order to avoid any ethical issues, full anonymisation is highly needed, though, it is worth mentioning here that Reddit provides an anonymised environment; which means that the author’s real name does not appear, but rather only the fake ones that have been chosen by the person when he/she signs up in this website. Reddit is an open-access public source for data. Proferes et al., (2021, p.3) illustrate that “Reddit posts, comments, and metadata can be accessed via the site itself, or via its APIs. Reddit’s official API is free and publicly available and provides an array of functions. However, the data retrieved from subreddits like r/SuicideWatch/r/depression is subject to the rules of these subreddits. The r/SuicideWatch subreddit, for example, prevents any use of its data without taking permission and revising the research by the moderators. So, the researchers have contacted the moderators, and permission has been taken to perform this study.

3.3. Compiling the corpus
As long as this study is corpus-based, the process of data collection should pass through two main steps; the first one is to compile a corpus; fifty suicide notes have been copied and stored electronically in “txt” file format as a self-build corpus (henceforth SNC) to deal with it easily using the corpus software and tools. Thus, the features of this corpus can be summarised as being specialised (established for a particular purpose in mind), written, monologue, and diachronic (from successive periods of time). Table 1 below shows the features of the study corpus.

While the second step is the annotation, which means marking each word in the corpus with its part of speech; this process is called POS tagging. The researchers have used the free TagAnt 1.2 software for POS tagging.

3.4. Words and sentences length
Examining the words and sentences length can be considered a good starting point to conduct a stylistic analysis of the selected data to identify its linguistic features.

This study employs WordSmith software to calculate the words and sentences in the corpus. Using the WordList tool, in the statistic section as named by the software, the words are classified into different categories depending on the number of letters each word includes. The sentences

| Table 1. Corpus information and abbreviations |
|---------------------------------------------|
| **Data**                                    |
| **Abbreviation** | **Corpus**   | **Size** | **Number of tokens** | **Year** |
| SNC         | Suicide notes corpus | 50 sample | 28,539               | 2012–2020 }
are organised by numbers to show how many sentences the corpus has; what can be considered as a sentence in this software is that any group of words starts with a word with a capital letter and ends with a full stop, exclamation, or question marks. Meanwhile, the sentence length is identified by counting the words each sentence includes.

According to Baddeley et al. (1975), in their experimental study, words that usually include 1–5 letters are considered short words, while the words that usually include 6–more letters are considered long words. Figure 5.1 below is a screenshot taken from WordSmith 8.0 software, illustrates the categorisation of the words that can be found in SNC:

In general, words are the essential components in any type of communication, whether written or spoken. So, the easier and shorter the words are, the simpler and uncomplicated the discourse will be; in contrast, the longer and more complicated the words are, the more complex and formal the discourse will be. From the above analysis, it becomes clear that the length of the words in the SNC ranges between 1–16 letters, and the words with 1,2,3,4 letters are the more frequent words in SNC.

Built on Baddeley et al., (1975) classification, the online suicide notes tend to include simple and short words more than long ones. Such results can be explained due to the following facts; first, the online suicide notes are usually presented to strangers to read, so the redditors in the SNC use simple and easy language to express themselves thoroughly. Second, and based on the findings of the previous studies like Baddeley et al (1975), shorter words are easy to remember and recall than longer ones. Concerning the mental status of the redditors in the SNC and the atmosphere of depression, sadness, and anger that prevails and controls them at the time of writing the letter. They usually recall the shorter words more efficiently and faster than the long complicated ones. Thus, the average word length in the suicide notes corpus is 3.84; This result confirms the information presented in Figures 2 and 3.

Concerning the sentence length, Sanyal (2006) has surveyed several USA participants to have a precise idea about the readability of the sentence, and what is regarded as an easy sentence to read and what is regarded as a difficult one. The result of such study indicates that: the sentence that usually includes eight words is “very easy” to read, while the sentence that includes eleven words is considered as an “easy” sentence to read, and the sentence that includes fourteen is considered as “a fairly easy” sentence to read, meanwhile the sentence with seventeen words is considered as a “standard” sentence, while the sentence with twenty-one words is considered as “fairly difficult,” and sentence with twenty-four words is considered as difficult. Lastly, a sentence that includes twenty-nine words is regarded as very difficult (p. 63). Regarding the length of the sentences in SNC, and by using WordSmith software, Figure 4. is presented:

Generally speaking, a text with shorter sentences is easily and concisely understood by readers because short sentences usually have a simple structure to convey the information directly and far away from being misleading. In contrast, a text with longer sentences is hard to be understood, and often a reader with inadequate knowledge would misinterpret it; longer sentences usually have a complex structure, rich details, and diverse contents. So, according to the information presented in Figure 4 above, and depending on the classification that Sanyal (2006) has presented in his study about the structure of sentences. The current researchers of this paper infer that the sentences in SNC are “fairly easy,” which means that the suicide notes language is not completely easy, nor it is completely difficult. It does include simple and easy words, but suicidal redditors’ desire to convey all what they want to say as their last words made them use relatively long sentences.

Figure 4 above also illustrates that the number of sentences in the SNC is very high compared to the number of samples that this corpus includes (50 suicide notes only). This conventionally means that the number of sentences in online suicide notes tends to be relatively large to convey
more information to the reader since they address public stranger readers, and the redditors try to express themselves entirely and precisely. Below are some examples of the sentences and words used in SNC:

Example 1:

“Hi. It is me. I have decided to end my life before my 28th birthday. Do not feel bad, this is what I have always wanted. ……….”

Example 2:

To my dear mother:

“I have decided to voluntarily take my own life. I am making this decision solely because I love you too much. I adore you to such a degree that I cannot bear to exist knowing that, by living, I risk the possibility of hurting you. ……….”

3.5. The lexical diversity

3.5.1. Token type ratio

Through using the Wordlist tool presented in WordSmith software, the type-token ratio of the words in each corpus can be identified easily. Token Type Ratio (TTR), according to Scott (2014), is the ratio between types and tokens in the text” (p. 398). The TTR can be generated using WordSmith 8.0 by counting the number of the tokens in the text (words) and the number of types in the text (only the different word types are counted, all the repeated ones should be deleted). It can be considered as a valuable analysis for equal-sized corpus only because it highly depends on the text size, and the result of such analysis between two corpora that vary in size would be inconsistent. However, to solve this problem, it is better to use Standardised TTR (STTR), which is used to present an average of Token Type ratio for each 1,000, and at the end, it presents an average of TTR.
Typically, the higher number of TTR of a text indicates that it is lexically diverse, while the lower number of TTR of a text indicates that the lexical in the texts is highly repetitive. The TTR of the SNC corpus is 10.28, while the STTR is 35.40%; This means that the suicide notes language is not lexically diverse, and the words used tend to be repetitive related to one or two themes. Generally speaking, from the results shown in Figure 5, the suicide notes corpus is not lexically diverse, and its words are highly repetitive. Such results show that the redditors in this corpus usually focus on one theme in their writings either blaming, thanking, advising, requesting, instructing others, just saying goodbye, or expressing their feelings. They do not mix more than one theme in their writings like any other, Redditors would do in everyday conversations/writings, which is why their vocabularies are highly repetitive.

3.6. Lexical distribution

According to McIntyre and Walker (2019), knowing the words and sentences length in a particular corpus alone would not be very useful in presenting a full consideration of the content of this particular corpus, but rather knowing the different word type and the most frequently repeated type is more efficient in identifying the stylistic features of a particular text. Furthermore, checking the lexical diversity and concentrating on the distribution of content words help reveal whether the text has nominal or verbal style (formal or informal), which is one of the significant characteristics of a text style that the stylistic analysis can check.

Concerning the current data, TagAnt software is used to mark up every word in the corpus with its part of speech; so the easiest way to examine the different parts of speech that is included in a corpus is to manage the data using the computer program Microsoft Excel to present a chart with the type of the words found and the frequency of each one in the current corpus under investigation. Figure 6. shows the types of words that are found in SNC and the frequency of each one:

After presenting a clear image of the lexical diversity in suicide notes language, it is essential to shed light on the distribution of content words, in particular, because that will help in characterising the text into a nominalise or verbalise text style to indicate whether the suicide notes are nominal preferences or verbal preferences. Moreover, since the pronouns are the most frequently used part of speech in suicide note corpus, it is helpful to examine the pronouns along with the other types of content words: nouns, verbs, adjectives, and adverbs.

3.7. Distribution of content words

Using both WordSimth software and Microsoft excel program, the annotated data (the data that have been dealt with by using TagAnt software to mark every word with its part of speech) have been sorted into different Wordlists; each one includes items that belong to a particular word class with its frequency of occurring in the study corpus. The distribution of the content words is identified by calculating the frequency of occurrence of each type of content word (nouns, verbs, adjectives, adverbs). The pronouns used in the SNC have also been examined.

As shown in Figure 7, in the SNC, the verbs are the most frequently used type of content word (6,603 freq.), sorting the verbs list by the frequency of its items, verbs like “be,” “have,” “do,” “want,” “know,” and “fell” get the highest numbers of frequency in the verbs list. While in the second place of the most frequent type of content word are the nouns with (4,305 freq.)
words like “life,” “people,” “time,” “suicide,” and “year” are the most frequent words in the nouns list. Furthermore, adverbs with (274 freq.) occur in the third place within the most frequently used type of the content words, words like “not,” “just,” “so,” “never,” “always,” and “really” are the most frequently used words in this list. At the end of the most frequently used content type list, adjectives occur (202 times). Adjectives like “good,” “sorry,” “happy,” “last,” and “much” are the most frequently used adjectives in the list. Finally, pronouns cover a high percentage of the tokens used in the SNC; they form 17.0% of the total number of its tokens, pronouns like “I,” “my,” “you,” “me,” and “it” get the highest numbers of frequency in the pronouns list.

From the above Figure 7, the verbs are the most frequent type to use in the SNC, then the pronouns and nouns cover a higher frequency, and regarding the adjective and adverbs both cover lower frequency in this corpus. So, generally speaking, and since the verbs are the most frequent content words in comparison to nouns and pronouns, one can confidently say that the suicide notes language is of verbal preference.

Redditors in the SNC usually employ informal style, using more verbs (because they usually want several actions to be done due to the fact that these are their last words) more pronouns (after checking the concordance line to the pronouns wordlist, the most frequent pronouns are I, my, you and me) more nouns (due to the fact they want to be direct and easy to be understood) and fewer numbers of adjectives (almost all the adjectives are negative ones) and adverbs.
3.8. The detailed analysis of the wordlists

For further information about each type of the content words used in this corpus, the researchers of this current study have analysed the Wordlists based on the selected categories that have been chosen from Leech and Short’s (2007) checklist. The results of such analysis indicate that:

- The verbs: Redditors in the SNC employ both transitive and non-transitive verbs in variant proportions, and these verbs carry an essential part in the meaning of the sentences. They also use non-factive verbs more frequently than the factive ones; verbs like feel, think, remember, thought, look, and wish are frequently used in their writings. Furthermore, they employ a narrative style of writing and focus more on describing their feelings by frequently using stative verbs.

- The nouns: Redditors in the SNC tend to use abstract nouns, which help in conceptualising the ideas and thoughts, more frequently than the concrete nouns; nouns like “life,” “time,” “years,” “nothing,” and “day” are used frequently by them. Furthermore, they frequently used proper nouns, and this can be explained due to the fact that they are more certain about their feelings for others, and their decisions are firm and straightforward. By mentioning the names, they do not allow themselves to withdraw or change their minds because they confess all their feelings to the people concerned by mentioning their names.

- The adverbs: the results of the analysis indicate that the redditors in the SNC employ different types of adverbs in different proportions. Adverbs are used to tell us more about the elements of the sentence, whether the verb, adjective, adverb, sentence, or paragraph. The most frequently used type of adverbs is the adverbs of times, followed by the additive/restrictive adverbs, then the linking adverbs, degree, manner, stance, and place adverbs, respectively.

- The adjectives: since the redditors in this corpus have used narrative style with simple and direct language, so they tend to use adjectives to decorate their writings to be more influential to the readers. They tend to use gradable adjectives rather than non-gradable ones and attributive adjectives rather than predictive ones. The types of adjectives used are arranged according to their frequency in the data as follows: general description adjectives, personality and emotion adjectives, emphasising adjectives, demonstrative adjectives, adjectives of age, and adjectives of degree.

- The pronouns in the SNC are extraordinarily frequent in this corpus; since the style of the texts is narrative style and the writers typically write about themselves, so the most frequently used pronoun is “I.” This pronoun represents 45.1% of the pronouns used and 7.7% of the total number of tokens in the SNC. Roughly speaking, the frequent use of the pronouns “I” refers to the need of the speaker to be seen as an individual, assertive, and the aware person responsible for his/her decision.

3.9. Keywords analysis

Keywords are markers or indicators of styles. To identify keywords in the SNC, the researchers have to extract a wordlist for each corpus (the Study corpus and the Reference corpus), and the software does all the calculation processes and presents a list of Keywords. WordSmith Keywords list includes both “positive words” and “negative words,” the words are classified based on the statistical differences in the frequency in each corpus. The words with a higher frequency in the study corpus are positive, and the words with a lower frequency are negative ones.

Concerning this study, the target/study corpus used is the SNC, which includes 28,548 tokens. As for the reference corpus used in this study, the researchers have selected an already existing dataset presented by Low et al. (2020). This corpus includes posts from 28 subreddits posted during 2018-2020. The number of its tokens is 156,945,984.

The problem with this reference corpus is that it includes posts from particular subreddits that might overlap with the current study data. Subreddit-like (r/Suicidewatch) (r/Suicidalthought) (r/Depression) have been removed from this corpus. What is left is a general picture of Reddit/sort of a corpus of general Reddit. The reason for choosing this reference corpus is that, as Culpeper
(2014) states, when both the reference corpus and the study corpus are close to each other, the result will be more focused and directed to the study corpus. In this study, the same genre is used (the corpora used contain online posts retrieved from the Reddit platform only).

The researchers need to set the statistical setting for the software to perform the analysis of the keyword and present a keywords list. The statistical setting used for this study was as follows: using the approximate Bayes factors as a measure for the effect size, at 6.000 degrees with the associate p-value 0.00001 and LL 15.13 to present strong evidence against the null hypothesis, which state that the difference in frequency of a particular item between the two corpora occurs by chance. Using Log-ratio as a mechanism to reduce the items in the keywords lists, any item whose log-ratio is 2 or above is considered a key. After setting the statistical tests required to produce the list of the keywords, a qualitative analysis of its items should be presented. Figure 8 shows a screenshot of the SNC keywords list used in this study:

At the top of the Keywords list are words that are completely in harmony with readers’ intuition about what the suicide notes are all about. Words like “you,” “life,” “sorry,” “Suicide,” “please,” “hope,” “happy,” “pain,” and “world” are the most frequently used words in this list. All these words are positive words, which means that the Redditors in the SNC use them more frequently than the redditors in the reference corpus.

The Word “you” is the most key in the keywords list, which can strongly support the intuition about suicide notes writers. They tend to instruct, blame, ease the pain, or show their love for their lovers. The second person pronoun “you” is usually used to refer to a particular person or people in general. In the current data, this pronoun is mainly used to refer to people in general. The redditors in the SNC address their surrounding people, either to put the blame on them for making this decision, as shown in the first 15 lines in Figure 9. These lines obviously belong to the same note when the writer/Redditor expresses his anger and blames others. The rest of the lines presented in Figure 9 illustrate that the Redditors in this corpus are trying to express their love and ease the pain from their lovers. This pronoun represents 2.06% of the words used in the SNC, while it represents 0.41% of the words used in the reference corpus. Figure 9. below shows the concordance lines of pronoun “you”.

Other keywords in the list can be seen as guessable keywords; typically, when writing a note, the writer ends this note by wishing something to the reader. A word like “hope,” which is at the top of this corpus’s Keywords list, is used to convey this meaning.

Another word at the top of the keywords list is “life,” which is typically the main topic in this type of writing. The reditors use this word frequently in the SNC when they talk about their own life; it represents 0.55% of the tokens in the SNC and 0.14% of the reference corpus tokens. The word “life” is spread all over the file, which becomes evident after checking the plot of its distribution, and this means that all the writers/redditors in the SNC have used this word while writing their notes. This word clusters with particular words, as shown in Figure 10:

The frequent use of other words in the keywords list like “please,” “Dear,” and “goodbye” are an indicator of the fact that the reditors in the SNC employ polite and respective language when they address their lovers, the word “please,” for example, represents 0.22% of the tokens used in the SNC, while in the Reference corpus it represents 0.04% of its tokens only. Figure 11 below shows that the frequent use of the word “please” is to ask polite requests.

As for the words “suicide,” “notes,” and “letters,” which can be seen among the frequently used words in the keywords list of the SNC. They are used frequently in this corpus because they must have existed within the title of each post. By using particular subreddits like (r/suicidewatch) or (r/ depression), the user is forced to present a title to his/her post, and since the main topic of this
A study is about suicide; so all the posts that are entitled “my suicide note,” “my suicide letter,” “suicide notes,” and the like have been collected and included within the data.

The remaining words in the keywords list represent the aboutness of this topic since this corpus includes posts about suicide, so typically, the words used to express feelings and emotions are to be seen at the top of the most frequently used words. These words can be classified into different groups; the first group includes words that are used to convey positive meaning. At the top of this group is the word “sorry” which is used frequently by redditors in the SNC to express their regrets about the wrong things that they have done or (they think they have done) toward their lovers.

Regarding the same group, the word “fault,” on the other hand, can also be included within this group; although it has been used to express negative feelings, the redditors in the SNC use this word to offer consolation to their lovers by stating directly, “it is my fault” or “it is not your fault,” as illustrated in Figure 12:

The same can be said concerning the word “blame,” which represents 0.05% of the words used in the SNC and no use in the reference corpus. The redditors use this word in the SNC to illustrate to their lovers that they do not blame anyone other than themselves and give them some condolences. Other words like “loved,” “happiness,” “strong,” “cared,” and “miss” are used to convey positive feelings for the lovers in this corpus.

The second group includes the words that the redditors have used to convey their negative feelings to their lovers. Words like “pain,” “failed,” “deserve,” “selfish,” “useless,” and “sadness” are used frequently by the writers when they are trying to justify their decisions and explain why they would commit such an act; the act of killing themselves.

While for the third group, includes the words that the redditors in the SNC have used to announce their intuitions to commit suicide. Words like “kill,” “die,” and “death” are the most frequently used words to reveal their desires to end their lives.
In the same vein, some of these words show a relatively lower frequency in the SNC, like the word “Apathy,” which is used 5 times in the corpus, “burden,” “escape,” and “deserved,” which are used 9 times only. But when sorting the words by their statistical significance, these words show a relatively higher degree of significance, and this can be explained due to the rare use of these words in the reference corpus by ordinary redditors. The redditors in the SNC use such words when they are trying to describe their mental states or themselves as in the way they regard themselves. They see themselves as a burden, apathetic person who deserves all what he/she has gone through.
3.10. Discussion

The study shows that redditors in the SNC use simple and easy language to express themselves thoroughly and short words due to the fact that the short words are easy to remember and recall. They use fairly easy sentences; in other words, the sentence in suicide notes are relatively short to help them convey their last words freely, though it is not long or completely long, but the suicide notes include large numbers of sentences compared with the number of samples included in the corpus (fifty suicide notes); the online suicide notes tend to be long. The online suicide notes include a lower type-token ratio, which means that the words in the corpus tend to belong to specific themes. Since the notes are not lexically diverse, the words are highly repetitive. Suicidal redditors focus in writing their notes on one thematic topic as illustrated in the given example two.

![Figure 11. The concordance lines of the word “please”](image)
I have decided to voluntarily take my own life. I am making this decision solely because I love you too much. I adore you to such a degree that I cannot bear to exist knowing that, by living, I risk the possibility of hurting you. ........

The writer of the note in this example thinks that she/he is doing the best choice of ending her/his life. As if committing this action is proof of love for the mother. This feeling of remorse is torturing the writer. Thus, she/he believes suicide is the protector of this unconditional love. Thus, the writing of suicidal people on Reddit tends to be long, repetitive, and lexically simple.

Regarding the distribution of content words in the online suicide notes, verbs are the most frequent type to use in the corpus, then the pronouns and nouns cover a higher frequency, which can confirm with the previous studies’ findings (e.g., Osgood & Walker’s, 1959; Roubidoux, 2012). Regarding the adjectives and adverbs, both cover lower frequencies in the corpus. Generally speaking, the suicide notes are verbal preferences rather than nominal preferences; since the Redditors in these kinds of posts usually ask the receivers to perform certain activities as their will before committing suicide.

As for the Keywords analysis, which aims to identify the aboutness of this corpus, the findings in this section reveal that Redditors in the SNC tend to frequently use emotional words, which help convey their positive and negative feelings to the readers.

3.11. Conclusion
This study is carried out to present a corpus-based stylistic analysis of a collection of online suicide notes retrieved from the online website Reddit to examine the stylistic and linguistic features of these notes. The analysis has been conducted based on a pre-prepared checklist of selected categories taken from Leech and Short (2007) checklist and with the help of corpus tools and software. The results of such analysis have emphasized the role of the corpus tools to aid the stylistic analysis; they
facilitate the analysis process and help to present accurate information concerning the quantitative analysis. The results demonstrate that the redditors in the online suicide corpus use simple words and short sentences. The reason for this inclination might guarantee reflecting ideas in a brief direct way, especially during stress and grief. Instead of using complicated long structured sentences that might hinder clear, straightforward meaning. Their writings tend to be less lexically diverse, concentrating on one or two themes with a lower TTR (type-token ratio), which indicates that their words were relatively repetitive. Moreover, the usage of first-person singular pronouns is quite frequent and common in the corpus of online suicide notes. Regarding the distribution of content words, verbs are the most frequently used part of speech, followed by nouns, adjectives, and adverbs. For the keywords analysis section, the findings reveal that they tend to use emotional words to express their intuition to end their lives more often than the ordinary redditors do in the reference corpus. The results also indicate that suicidal redditors address their family members or beloved ones with polite considerate language.

Funding
The authors received no direct funding for this research.

Author details
Eman Adil Jaafar
E-mail: eman_jafer@coeduw.uobaghdad.edu.iq
ORCID ID: http://orcid.org/0000-0002-8880-8930

Haya Abdul-Salam Jasim
1 Department of English, College of Education for Women, University of Baghdad, Baghdad, Iraq.

Disclosure statement
No potential conflict of interest was reported by the author(s).

Corrections
This article has been republished with minor changes. These changes do not impact the academic content of the article.

Citation information
Cite this article as: A corpus-based stylistic analysis of online suicide notes retrieved from Reddit, Eman Adil Jaafar & Haya Abdul-Salam Jasim, Cogent Arts & Humanities (2022), 9: 2047434.

References
Anthony, L. (2015). TagAnt [Computer Software] (1.2.0). Waseda University. https://www.laurenceanthony.net/software/tagant/

Boddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and short-term memory. Journal of Verbal Learning and Verbal Behavior, 14(6), 575–589. https://doi.org/10.1016/S0022-5371(75)80045-4

Barak, A., & Miron, O. (2005). Writing characteristics of suicidal people on the internet: A psychological investigation of emerging social environments. Suicide and Life-Threatening Behavior, 35(5), 507–524. https://doi.org/10.1521/sult.2005.35.5.507

Burnap, P., Colombo, G., & Scourfield, J. (2015). Machine classification and analysis of suicide-related communication on twitter. Proceedings of the 26th ACM Conference on Hypertext & Social Media (US: Association for Computing Machinery), 75–84. https://doi.org/10.1145/2700171.2791023

Collins Concise English Dictionary. (2014). Suicidal. In Collins concise english dictionary. HarperCollins. Retrieved December 15, 2020, from https://www.colinsdictionary.com

Culpeper, J. (2014). Keywords and characterization an analysis of six characters in Romeo and Juliet. In D. L. Hoover, J. Culpeper, & K. O’Halloran (Eds.), Digital literary studies: Corpus approaches to poetry, prose, and drama (pp. 9–35). Routledge.

Edelman, A. M., & Renshaw, S. L. (1982). Genuine versus simulated suicide notes: An issue revisited through discourse analysis. Suicide and Life-Threatening Behavior, 12(2), 103–113. https://doi.org/10.1111/j.1943-278X.1982.tb00917.x

Greaves, M., & Dykeman, C. (2019). A corpus linguistic analysis of public Reddit blog posts on non-suicidal self-injury. arXiv preprint arXiv:1902.06689. https://arxiv.org/abs/1902.06689

Gregory, A. (2018). The decision to die: The psychology of the suicide note. In D. Cantor, and L. Allison (Eds.), Interviewing and deception. Routledge (pp. 127–156).

Jasim, H. A., & Jaafar, E. A. (2022, January-March). Studies on linguistic stylistic analysis of suicide notes and suicidal thoughts posts. International Journal of Research in Social Sciences & Humanities, 12(1), 100–124. https://www.jrssh.com/admin/upload/06%20Haya%20Abdul%2001289.pdf

Jones, C., & Woller, D. (2015). Corpus linguistics for grammar: A guide for research. In Corpus linguistics for grammar. Routledge. https://doi.org/10.4324/9781315171779

Leech, G., & Short, M. (2007). Style in Fiction. Pearson Education Limited. https://doi.org/10.4324/9781315835525

Low, D. M., Rumker, L., Tolkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (2020). Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study. Journal of Medical Internet Research, 22(10), e22635. https://doi.org/10.2196/22635

McIntyre, D., & Walker, B. (2019). Corpus stylistics: Theory and practices. Edinburgh University Press.

Osgood, C. E., & Walker, E. G. (1959). Motivation and language behavior: A content analysis of suicide notes. Journal of Abnormal and Social Psychology, 59(2), 58–67. https://doi.org/10.1037/h0047078

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A systematic overview of disciplines, approaches, methods, and ethics. Social Media+ Society, 7(2). https://doi.org/10.1177/20563051211033823

Roubloux, S. M. (2012). Linguistic manifestations of power in suicide notes: An investigation of personal pronouns. University of Wisconsin Oshkosh.

Sanyal, J. (2006). Indilish: The book for every english-speaking Indian. Vive Books Ltd.

Scott, M. (2014). WordSmith tools help. Lexical Analysis Software.

Scott, M. (2021). WordSmith tools [Computer software]. Lexical Analysis Software. https://www.lexically.net/wordsmith/downloads/

World Health Organization. (n.d.). Suicide. (2019 Sep 2). Retrieved December 28, 2020, from https://www.who.int/news-room/fact-sheets/detail/suicide
