Computational Analysis of the Historical Changes in Poetry and Prose

Amitha Gopidi
IIIT, Hyderabad
amitha.g@research.iiit.ac.in

Aniket Alam
IIIT, Hyderabad
aniket.alam@iiit.ac.in

Abstract

The esoteric definitions of poetry are insufficient in enveloping the changes in poetry that the age of mechanical reproduction has witnessed with the widespread proliferation of the use of digital media and artificial intelligence. They are also insufficient in distinguishing between prose and poetry, as the content of both prose and poetry can be poetic. Using quotes as prose considering their poetic, context-free and celebrated nature, stylistic differences between poetry and prose are delved into. Grammar and meter are justified as distinguishing features. Datasets of popular prose and poetry spanning across 1870-1920 and 1970-2019 have been created, and multiple experiments have been conducted to prove that prose and poetry in the latter period are more alike than they were in the former. The accuracy of classification of poetry and prose of 1970-2019 is significantly lesser than that of 1870-1920, thereby proving the convergence of poetry and prose.

1 Introduction

Language is the mathematics of expression. It is a mathematics because one stitches together an algorithm of concepts in the world, that we identify through words. This world of words is akin to dealing with numbers, because both in their atomic or denotative sense convey very little. But when they combine, they have unlimited potential of justifying profound concepts of time and space.

In this mathematical world of language, however, the origin and definitions of poetry as put forth by philosophers are esoteric in nature with little verifiability. And most of these esoteric definitions, though exotic, can be applied just as well to prose. Plato, in Republic, Book X, writes that poetry has the power to transform its audience, and poets therefore should be held accountable for what they write given this transformative power.

Aristotle (Golden, 1968), differentiates between different artforms but his discussion of poetry as being mainly tragedy or comedy explains that poetry since then has become a much wider artform. Kant (1952) expounded that a "poem may be very neat and elegant, but without spirit" if it lacks imagination, while the same may be said about prose. Shelley (2009) insists that poets are the "unacknowledged legislators of the World" and poetry to him is the "expression of the imagination" which he opines comes naturally to mankind. He also gives a restricted definition of poetry as follows:

> Poetry in a mere restricted sense expresses those arrangements of language, and especially metrical language, which are created by that imperial faculty, whose throne is curtained within the invisible nature of man.

However, as pointed out by Gioia (2003), poetry is a rapidly changing art. He writes that "the general term poetry, for example, now encompasses so many diverse and often irreconcilable artistic enterprises that it often proves insufficient to distinguish the critical issues at stake." Admittedly then, the definition of a poem in the state of the situation is the poem itself. And this causes a problem because anything goes in the name of poetry.

This paper is an attempt to understand what has changed in poetry over the last 150 years within the age of mechanical reproduction of art, named so by Benjamin and Underwood (1998). Comparison, therefore, has been done between the early stages of this age, wherein romantic poetry flourished, namely 1870-1920 and the late stage, 1970-2019, which saw the mechanical reproduction of art occurring through various digital forms. The latter importantly saw the creation of artworks

using artificial intelligence, with many computational poetry generators spewing poetry.

> Whatever one thinks of the artistic quality of the new poetic forms, one must concede that at the very least they reassuringly demonstrate the abiding human need for poetry.
>
> (Dana Gioia)

1.1 Prose and Poetry - Differentiating features

In the attempt to learn how poetry has changed over the last 150 years, the features that are normally attributed to poetry were studied. However, it was noticed that semantic features such as imagery, metaphors, sentiment, choice of words, themes, topics and associations were not strictly ascribed to poetry. All of these features can also be found in prose, and it is for this reason that prose is also called 'poetic', as corroborated by Eagleton (2007). Toni Morrison, for instance, is called a highly 'poetic' writer (Beaulieu, 2003). While we see that many works of prose are poetic, choosing entire novels would cause a lot of noise in the data. It is for this reason that we carefully chose quotes from popular novels as our prose, because quotes are the touchstones of books, are contextually independent of the situation in the book and hence make sense in a stand alone manner.

The visual difference between a quote and a poem are the line breaks.

> A poem is a fictional, verbally inventive moral statement in which it is the author, rather than the printer or word processor, who decides where the lines should end.
>
> (Terry Eagleton)

This, however, is also to say that a quote can be converted to a poem by an individual's decisions as to where to split the sentences into new lines. For this reason, line breaks were avoided as a feature.

Grammar, however, was identified as an important differentiator between poetry and prose by the authors by manual evaluation of the prose and poetry datasets. Within grammar, different types of inversions of word orders in sentences such as verb-subject inversion, along with dependent clauses, questions and conjunctions were chosen as features and are justified under section 2.2.

Meter was also considered as a feature, because as explained by Boulton (2014), meter is only a subsection of rhythm, and meter consists of the most identifiable rhythms. It is also important to note that she makes it abundantly clear that "free verse is not some glorious revolutionary emancipation of poetry, allowing sincerities never before possible." But that it is the kind of poetry with a meter that is neither traditional nor recognizable. And inversions are mentioned to be used in order to enforce a metrical structure in a poem, and therefore, it made all the more sense to consider meter as a feature.

Rhyme, however, was only used in comparing poetry of the two chosen periods and is also aided by inversion.

1.2 Related Work

Classification approaches between poetry and prose have been done by Roxas and Tapang (2010) using word adjacency networks and latent dirichlet allocation. Jamal et al. (2012) have attempted a classification of just poetry using themes. Tanasescu et al. (2016) have done a classification of poetry with respect to only rhyme and meter.

In work related to analysis of prose and poetry, Doumit et al. (2013) have worked on differentiating prose and poetry of two popular poets and authors each, using a semantic neural model. They show that poetry possesses a higher number of associations than prose. However, quotes are as poetic as prose with many metaphors and associations. Therefore, our problem is unique as we differentiate between quotes and poetry.

Kao and Jurafsky (2015) have done a computational analysis of poetic style using amateur and professional poetry. However, they concentrate on parts-of-speech tag occurrences and semantic features such as imagery, emotional language, sound devices and diction. Semantic features in this paper have been avoided with a rationale that quotes and poetry would be very similar with respect to these.

Chen et al. (2014) have worked on converting prose into rhyming verse, which uses substitution choices so as to enforce rhyme and produces sonnets based on an input of source sentences.

Computational poetry generators have used prose in the form of input or as training data, applying constraints on it using meter, rhyme and type of words through deep learning as

well as heuristic approaches (Chen et al., 2014; Ghazvininejad et al., 2016; Yi et al., 2017).

Therefore, the use of different sentences styles in poetry as compared to prose contributed by the use of inversion has not been used as a feature so far. While meter was widely used, the dynamics between meter, inversion and rhyme have not been explored. Our high accuracies of classification with just inversion as a feature show the importance of the account of sentences-styling in poetry as compared to prose. The study of change in poetry with respect to prose historically is also unique to this paper, clearly showing the dwindling of the features that were once more prevalent in poetry than they are today. The absence of change in prose over the years with respect to the stylistic features of inversion and meter is also shown.

2 Methodology

2.1 Dataset Overview

The four datasets using which our features for the historical analysis of poetry and prose were derived, belong to two time segments 1870-1920 and 1970-2019. The reasoning behind choosing these particular time segments is explained in Section 1. Each time segment has a dataset of both prose and poetry. Each of the datasets were made computationally by curating content written in the respective time segment by popular poets and books of the time.

For poetry, PoetryFoundation[1] and PoemHunter[2] websites were used. Finding the year of publishing of individual poems was difficult, so lists of popular poets of that time period were manually chosen from the websites mentioned, and their works were collated in the form of pdf files. These pdf files were converted into datasets.

For prose, 30 top liked quotes (or lesser if 30 weren't available) from 500 most popular books of the time segment as listed by Goodreads[3] were computationally collected. The top liked quotes are often quite 'poetic' in their content. The meta structure of our datasets is described in table 1.

2.2 Features

Each line of a poem, and each sentence of a quote was considered as the smallest unit on which the

[1]https://www.poetryfoundation.org/
[2]https://www.poemhunter.com/
[3]https://www.goodreads.com/

Type	Time Period	Count
Prose	1870-1920	7838 quotes
Prose	1970-2019	12623 quotes
Poetry	1870-1920	13635 poems
Poetry	1970-2019	7917 poems

Table 1: Datasets

following features were calculated:

2.2.1 Grammar

While prose is always grammatical, poetry tends to break away from the limitations of grammar. With regard to the celebrated poet Emily Dickinson, Miller (1987) writes that the former often wrote in an ungrammatical manner. The term 'poetic license' (Britannica, 2007) is a testament to the fact that poets often break the rules of grammar. For instance, Kaur (2017) and Cummings (1994) have written without capitalization or punctuation, thus violating grammar.

While the lack of capitalization and punctuation are not universal among poems, by manual evaluation, it was noticed that the styles of the sentences used in poetry greatly differed from those in prose because of the use of inversion. Inversion is defined as, "the syntactic reversal of the normal order of the words and phrases in a sentence, as, in English, the placing of an adjective after the noun it modifies ("the form divine"), a verb before its subject ("Came the dawn"), or a noun preceding its preposition "worlds between")" (Britannica, 2016).

As an example, Wordsworth in his poem, "I Wandered Lonely As A Crowd" (Wordsworth) uses the verb-subject inversion when he writes "Ten thousand saw I at a glance" instead of "I saw ten thousand at a glance".

We use four different kind of inversions that we observed in poetry based on the discussion of styling sentences in Waddell (1993) supplemented by the insights in Literary Devices website (Devices, 2015).

Along with these, features such as dependent clause as a subject, rhetorical questions and lines beginning with conjunctions are used as features. The use of conjunctions at the beginning of a line/sentence is disputed to be ungrammatical (Soanes, 2012), but we noticed that the usage was higher in poetry as compared to prose considering that poetry is a grouping of phrases and clauses.

16

Waddell (1993) also describes the use of dependent clauses as a pattern of styling sentences and we noted that dependent clause as a subject occurred quite often in poems. The use of rhetorical questions in literature (Devices, 2017) is quite prevalent, and they occurred more in our poetry datasets. The list of features related to grammar with examples are listed in table 2.

In order to implement all of the above features, Stanford CoreNLP (Manning et al., 2014) tools of tokenization, parts of speech tags, dependency parse trees, OpenIE triples were used. Simple heuristics were used to decide which kind of inversions exist in a given sentence using POS tags and OpenIE triple occurrences in the sentence. For instance, if the subject given by OpenIE triple of a line is a noun or pronoun, and it is preceded by a verb, the line would be marked as having subject-verb inversion. The OpenIE tool trained on prose, doesn't always fetch results for lines in poetry and in these cases, we use POS tags as they are accurate for poetry as well. The various inversion counts were normalized by the number of lines in poetry datasets and the number of sentences in prose datasets, so as to remove dependency on the the length of the poem or quote.

2.2.2 Meter

Poets use inversion in order to fit their material into a meter (Britannica, 2016), which is nothing but the arrangement of stressed(s) and unstressed(w) syllables in a certain manner (Boulton, 2014). In order to implement meter, we used Stanford Literary Lab's Poesy (Heuser et al., 2018), which is a python module for poetic processing. The module gives information of a base meter among four types of base meters:

1. Iambic [ws]

2. Trochaic [sw]

3. Anapestic [wws]

4. Dactylic [sww]

It also gives information regarding the number of repetitions of meter in a given line, thus leading to information on whether the poem is a pentameter, hexameter etc.

2.2.3 Rhyme

Inversion is also used often to fit into a rhyme scheme along with meter. Using the Poesy

(Heuser et al., 2018) module, we also extract the rhyme scheme of poems. It was only used for comparison between the two poetry datasets. It has not been applied on the prose datasets as the values were null.

2.3 Classification

The feature vectors consisted of 9 features discussed in the previous sections. Seven of them are various inversion types, followed by the base meter and number of feet. The extra feature, 'rhyme type' was only used for classification between the two poetry datasets.

The feature data was trained through a random forest classifier (Breiman, 2001) and KNN classifier with a 70/30 split for the training and testing data. The optimal value of the number of trees for random forest classifier was found to be 100. The value of k is taken be 3 for the KNN classifier. To deal with class imbalance, we adjust weights inversely proportional to class frequencies in the data.

Four experiments of different classifications between poetry and prose were conducted.

3 Results

Random Forest classifier performed better than KNN classifier in all of the below experiments:

3.1 Prose vs Poetry of Each Period

Classification of prose and poetry of each period was done to see if classification accuracy between poetry and prose has reduced for the time segment 1970-2019 as compared to that of 1870-1920. This would indicate that poetry and prose are more similar in 1970-2019 than they were in 1870-1920. Various combinations of features were used with both the classifiers.

The reduction in the classification accuracy of poetry and prose of 1970-2019 as shown in table 4 as compared to 1870-1920 as shown in table 3, indicates convergence in poetry and prose in the period 1970-2019.

3.2 Poetry: 1870-1920 vs 1970-2019

Classification of poetry of 1870-1920 and poetry of 1970-2019 was conducted with rhyme as an additional feature.

The results as shown in table 5 indicate that poetry has undergone a significant change with an accuracy of 77% in classification.

Feature	Example
Adjective Inversion: Adjective occurs right after the noun.	"I sing the body electric"
Subject Verb Inversion: Verb occurs before its subject.	"Ten thousand saw I at a glance."
Prepositional Phrase Inversion: Prepositional phrase occurs before subject and verb, or verb and subject.	"Until we meet again, to be counted as bliss."
The Yoda construction: Modifier followed by subject and verb.	"Whose woods these are I think I know."
Dependent Clause as a subject, followed by verb.	"What man cannot imagine, he cannot create."
Question	"Shall I compare thee to a summers day?"
Beginning with a conjunction	"Two roads diverged in a yellow wood, And sorry I could not travel both And be one traveler, long I stood And looked down one as far as I could To where it bent in the undergrowth;"

Table 2: Grammatical Features

Classifier	Feature	Accuracy	F1 Score	ROC AUC Score
Random Forest	All	98.4	98.4	98.02
Random Forest	Meter	93.7	93.7	93.8
Random Forest	Inversion	91.8	91.8	91.3
kNN	All	97.4	97.4	97.02
kNN	Meter	93.6	93.6	93.8
kNN	Inversion	91.8	91.8	91.2

Table 3: Poetry vs Prose classification results for 1870-1920

Classifier	Feature	Accuracy	F1 Score	ROC AUC Score
Random Forest	All	91.5	91.3	89.7
Random Forest	Meter	83.4	82.8	80.06
Random Forest	Inversion	85.1	84.5	82.06
kNN	All	90.0	89.8	88.2
kNN	Meter	80.6	79.2	75
kNN	Inversion	84.1	83.6	81.07

Table 4: Poetry vs Prose classification results for 1970-2019

Classifier	Accuracy	F1 Score	ROC AUC Score
Random Forest	77.0	76.3	73.11
kNN	72.6	72.5	70

Table 5: Poetry 1870-1920 vs Poetry 1970-2019 classification

Classifier	Accuracy	F1 Score	ROC AUC Score
Random Forest	59.2	50.7	50
kNN	51.4	52.09	51.1

Table 6: Prose 1870-1920 vs Prose 1970-2019 classification

Classifier	Accuracy	F1 Score	ROC AUC Score
Random Forest	94.7	94.7	94.7
kNN	94.04	94.0	94.01

Table 7: Poetry vs Prose classification overall

3.3 Prose: 1870-1920 vs 1970-2019

Classification of prose of 1870-1920 and prose of 1970-2019 was conducted.

The results as shown in table 6 indicate that nothing much has changed in prose as per these features over the two periods because the evaluation scores are close to a random guess (59%).

3.4 Poetry and Prose Both Periods Combined

This classification was done with combined datasets of poetry against combined datasets of prose. As per the results shown in table 7, given an input, this classifier would differentiate between it being a poem or prose with 94.7% accuracy. This is an important result as we do not consider line breaks.
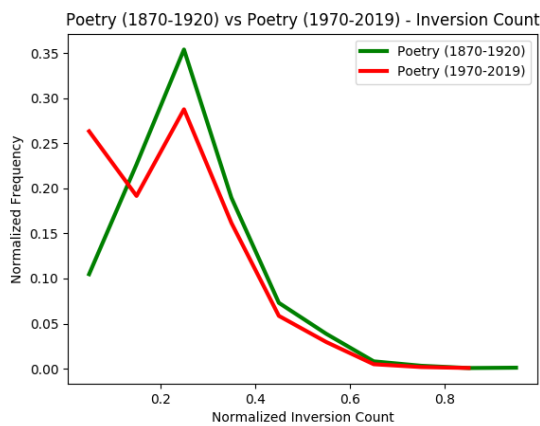
Figure 1: Poetry (1870-1920) vs Poetry (1970-2019) - Inversion Count

4 Analysis

4.1 Inversion

Figures 1 and 2 are plotted between the normalized inversion count(so as to remove any dependency on the length of the poem/prose), and the normalized frequency of the datasets(so as to remove dependency on the number of data points).
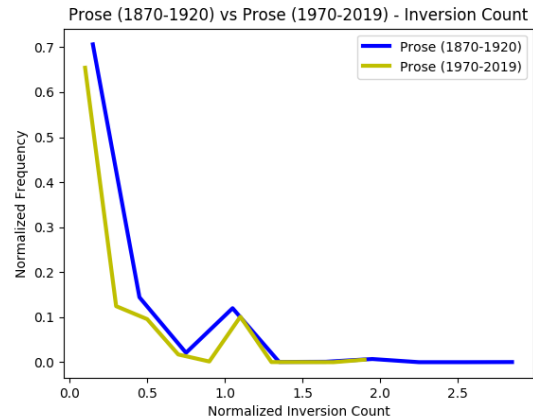

Figure 2: Prose (1870-1920) vs Prose (1970-2019) - Inversion Count

Figure 1 indicates a significant fall in the inversion count in the second time period. And figure 2 shows that the inversion counts of the two periods of prose are more or less the same.

4.2 Meter Base Type

The figures 3 and 4 represent the historical change in meter over the two periods for both prose and poetry. The y axis represents the percentage of dataset, which is a normalized indicator and does not skew the graph towards the period with higher data points.
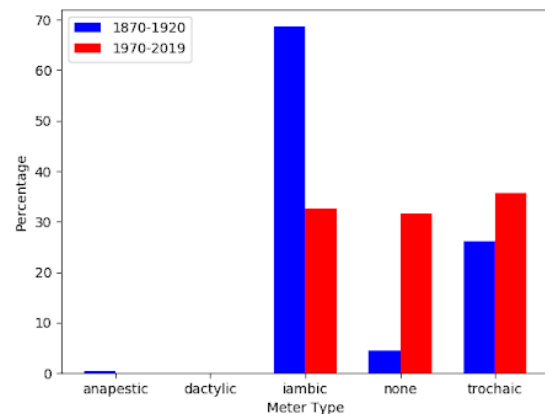

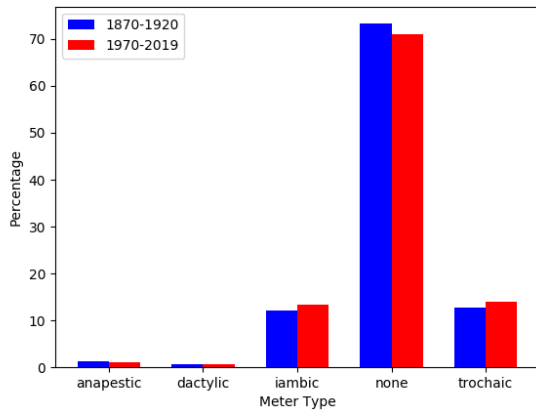Figure 3: Poetry (1870-1920) vs Poetry (1970-2019) - Meter Type

19

Figure 4: Prose (1870-1920) vs Prose (1970-2019) - Meter Type

Figure 3 clearly shows the dominance of the iambic base meter in poetry datasets, and its fall from 1870-1920 to 1970-2019. It also shows that in 1970-2019, the number of poems with no distinguishable meter has risen considerably with no significant change in anapestic or dactylic base meters.

Figure 4 proves that there is no significant difference in the base meter of prose over the two chosen periods, with none value as the most dominant.

4.3 Popular Meters

Figure 5 and Figure 6 were drawn so as to show the change in meter. Meter plotted is a combination of the base meter and its feet in the poem. The top 7 meters were chosen for the plots.

Figure 5 shows the significant fall of the all the popular meters in the second time period as compared to the first. The increase in the none values also suggests that the second time period consists of poetry with no recognizable meter.
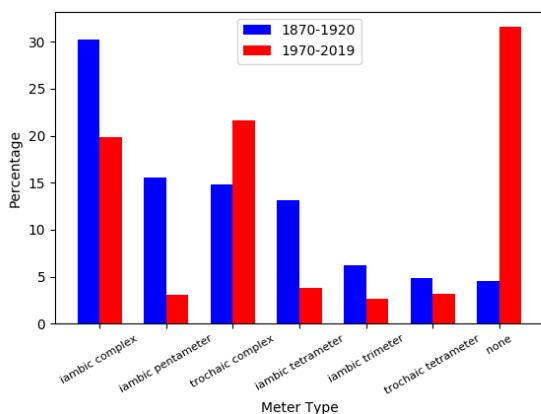

Figure 5: Poetry (1870-1920) vs Poetry (1970-2019)

Figure 6, on the other hand, shows the lack of recognizable meter in prose which is expected.
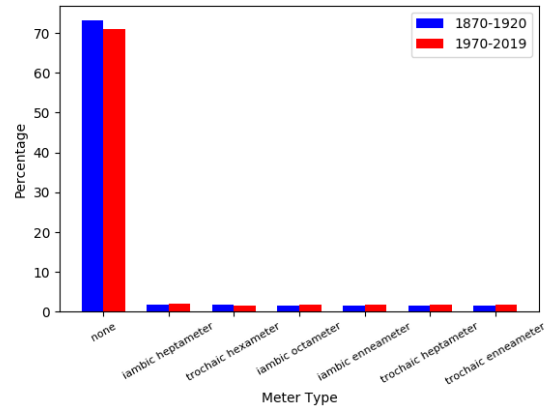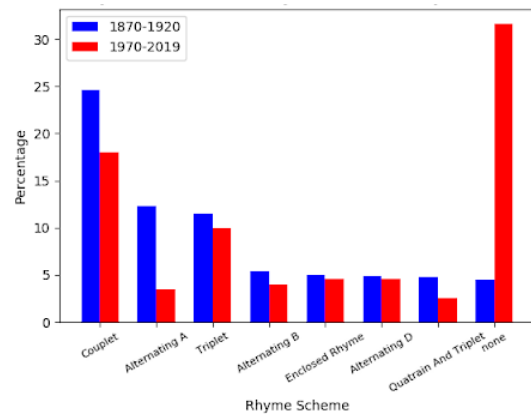

Figure 6: Prose (1870-1920) vs Prose (1970-2019)

Figure 7: Poetry (1870-1920) vs Poetry (1970-2019) - Rhyme Scheme

4.4 Rhyme

The rhyme feature used in figure 7, shows that a large percentage of 1970-2019 poetry has no rhyme scheme, while also showing that the prevalence of the other rhyme schemes has also come down.

5 Conclusion

From the experiments conducted, it has been proved that the poetry of 1970-2019 is more similar to prose of its period than the poetry of 1870-1920 was to the prose of the same period. The changes in prose of the two periods with respect to stylistic features are minimal, but those in poetry are significant. The convergence of poetry and prose and lack of change in prose, proves that poetry does not possess the liminal boundaries that prose enjoys. The importance of a new age defi-

20

nition of poetry is thus established considering the changes in poetry as an artform.

Apart from justifying the historical changes in poetry and prose, this paper also achieves high accuracy in the classification of poem and prose using no semantic features. This is an important indicator that semantic content of poetry and prose can be very alike and that they can still be differentiated using stylistic features without considering the obvious visual difference of line breaks.

The future work of this paper is to use these features in constructing a personalized poetry assistant that learns the stylistic preferences of the user in inversions and meter, based on user input of creative text. This personalized nature of the assistant would adapt to the user's wishes in becoming a 'modern' or a 'classic' poet.

References

Elizabeth Ann Beaulieu. 2003. *The Toni Morrison Encyclopedia*. Greenwood Publishing Group.

Walter Benjamin and JA Underwood. 1998. *The work of art in the age of mechanical reproduction*. na.

Marjorie Boulton. 2014. *The Anatomy of Poetry (Routledge Revivals)*. Routledge.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

The Editors of Encyclopaedia Britannica. 2007. Poetic license.

The Editors of Encyclopaedia Britannica. 2016. Inversion.

Quanze Chen, Chenyang Lei, Wei Xu, Ellie Pavlick, and Chris Callison-Burch. 2014. Poetry of the crowd: A human computation algorithm to convert prose into rhyming verse. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

Edward Estlin Cummings and George James Firmage. 1994. *ee cummings: Complete Poems, 1904-1962*. Liveright Publishing Corporation.

Literary Devices. 2015. Inversion Examples and Definition.

Literary Devices. 2017. Rhetorical question.

Sarjoun Doumit, Nagendra Marupaka, and Ali A Minai. 2013. Thinking in prose and poetry: A semantic neural model. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Terry Eagleton. 2007. *How to read a poem*. John Wiley & Sons.

Marjan Ghazvininejad, Xing Shi, Yejin Choi, and Kevin Knight. 2016. Generating topical poetry. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1183–1191.

Dana Gioia. 2003. Disappearing ink: Poetry at the end of print culture. *The Hudson Review*, 56(1):21–49.

Leon Golden. 1968. *Aristotle's Poetics: a translation and commentary for students of literature*. Prentice-Hall.

Ryan Heuser, Mark Algee-Hewitt, Maria Kraxenberger, J.D. Porter, Jonny Sensenbaugh, and Justin Tackett. 2018. Poesy.

Noraini Jamal, Masnizah Mohd, and Shahrul Azman Noah. 2012. Poetry classification using support vector machines. *Journal of Computer Science*, 8(9):1441.

Immanuel Kant, Immanuel Kant, and Sir James Creed MEREDITH. 1952. *The Critique of Judgement. Translated with Analytical Indexes by James Creed Meredith*. Oxford.

Justine T Kao and Dan Jurafsky. 2015. A computational analysis of poetic style. *LiLT (Linguistic Issues in Language Technology)*, 12.

Rupi Kaur. 2017. *The sun and her flowers*. Simon and Schuster.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Cristanne Miller. 1987. *Emily Dickinson: A Poet's Grammar*. Harvard University Press.

Ranzivelle Marianne Roxas and Giovanni Tapang. 2010. Prose and poetry classification and boundary detection using word adjacency network analysis. *International Journal of Modern Physics C*, 21(04):503–512.

Percy Bysshe Shelley. 2009. A defence of poetry by percy bysshe shelley.

Catherine Soanes. 2012. Can you start a sentence with a conjunction?

Chris Tanasescu, Bryan Paget, and Diana Inkpen. 2016. Automatic classification of poetry by meter and rhyme. In *The Twenty-Ninth International Flairs Conference*.

Marie L Waddell. 1993. *The art of styling sentences: 20 patterns for success*. Barron's Educational Series.

William Wordsworth. I wandered lonely as a cloud by william wordsworth.

Xiaoyuan Yi, Ruoyu Li, and Maosong Sun. 2017. Generating chinese classical poems with rnn encoder-decoder. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 211–223. Springer.