frontiers
in Digital Humanities

ORIGINAL RESEARCH
published: 05 April 2018
doi: 10.3389/fdigh.2018.00005

The Gutenberg English Poetry Corpus: Exemplary Quantitative Narrative Analyses

Arthur M. Jacobs[1,2,3]*

[1] *Department of Experimental and Neurocognitive Psychology, Freie Universität Berlin, Berlin, Germany,* [2] *Dahlem Institute for Neuroimaging of Emotion (D.I.N.E.), Berlin, Germany,* [3] *Center for Cognitive Neuroscience Berlin (CCNB), Berlin, Germany*

This paper describes a corpus of about 3,000 English literary texts with about 250 million words extracted from the Gutenberg project that span a range of genres from both fiction and non-fiction written by more than 130 authors (e.g., Darwin, Dickens, Shakespeare). Quantitative narrative analysis (QNA) is used to explore a cleaned subcorpus, the *Gutenberg English Poetry Corpus* (GEPC), which comprises over 100 poetic texts with around two million words from about 50 authors (e.g., Keats, Joyce, Wordsworth). Some exemplary QNA studies show author similarities based on latent semantic analysis, significant topics for each author or various text-analytic metrics for George Eliot's poem "How Lisa Loved the King" and James Joyce's "Chamber Music," concerning, e.g., lexical diversity or sentiment analysis. The GEPC is particularly suited for research in Digital Humanities, Computational Stylistics, or Neurocognitive Poetics, e.g., as training and test corpus for stimulus development and control in empirical studies.

Keywords: quantitative narrative analysis, digital literary studies, neurocognitive poetics, culturomics, language model, neuroaesthetics, affective-aesthetic processes, literary reading

OPEN ACCESS

Edited by:
Robert J. Morrissey,
University of Chicago, United States

Reviewed by:
Charles Cooney,
University of Chicago, United States
Jan Rybicki,
Jagiellonian University, Poland

****Correspondence:***
Arthur M. Jacobs
ajacobs@zedat.fu-berlin.de

Specialty section:
This article was submitted to
Digital Literary Studies,
a section of the journal
Frontiers in Digital Humanities

Received: *24 November 2017*
Accepted: *08 March 2018*
Published: *05 April 2018*

Citation:
Jacobs AM (2018) The Gutenberg English Poetry Corpus: Exemplary Quantitative Narrative Analyses. Front. Digit. Humanit. 5:5. doi: 10.3389/fdigh.2018.00005

INTRODUCTION

In his "The psycho-biology of language," Zipf (1932) introduced the law of linguistic change claiming that as the frequency of phonemes or of linguistic forms increases, their magnitude decreases. Zipf's law elegantly expresses a tendency in languages to maintain an equilibrium between unit length and frequency, suggesting an underlying law of economy. Thus, Zipf speculated that humans strive to maintain an emotional equilibrium between *variety* and *repetitiveness* of environmental factors and behavior and that a speaker's discourse must represent a compromise between variety and repetitiveness adapted to the hearer's tolerable limits of change in maintaining emotional equilibrium. In a way, Zipf not only was a precursor of contemporary natural language processing/NLP (e.g., Natural Language Tool Kit/NLTK; Bird et al., 2009), quantitative narrative analysis (QNA), Computational Linguistics or Digital Humanities, but also of Psycholinguistics and Empirical Studies of Literature, since he theorized about "the hearers responses" to literature.

About 30 years later, when analyzing Baudelaires poem "Les chats," Jakobson and Lévi-Strauss (1962) counted text features like the number of nasals, dental fricatives, liquid phonemes or adjectives, and homonymic rhymes in different parts of the sonnet (e.g., the first quatrain) to support their qualitative analyses and interpretation of, e.g., oxymora that link stanzas, of the relation between the images of cats and women, or of the poem as an open system which progresses dynamically from the quatrain to the couplet. While their systematic structuralist pattern analysis of a poem starting with formal metric, phonological, and syntactic features to prepare the final semantic analysis provoked a controversy among literary scholars, it also settled the ground for subsequent linguistic perspectives on the analysis (and reception) of literary texts called *cognitive poetics* (e.g., Leech, 1969; Tsur, 1983; Turner and Poeppel, 1983; Stockwell, 2002).

Today, technological progress has produced *culturomics*, i.e., computational analyses of huge text corpora (5,195,769 digitized books containing ~4% of all books ever published) enabling researchers to observe cultural trends and subject them to quantitative investigation (Michel et al., 2011). More particularly, *Digital Literary Studies* now "propose systematic and technologically equipped methodologies in activities where, for centuries, intuition and intelligent handling had played a predominant role" (Moretti, 2005; Ganascia, 2015).

One promising application of these techniques is in the emerging field of *Neurocognitive Poetics* which is characterized by neurocognitive (experimental) and computational research on the reception of more natural and ecologically valid stimuli focusing on *literary materials*, e.g., excerpts from novels or poems (Schrott and Jacobs, 2011; Jacobs, 2015a,b; Willems and Jacobs, 2016). These present a number of theoretical and methodological challenges (Jacobs and Willems, 2018) regarding experimental designs for behavioral and neurocognitive studies which—on the stimulus side—can be tackled by using advanced techniques of NLP, QNA, and machine learning (e.g., Mitchell, 1997; Pedregosa et al., 2011; Jacobs et al., 2016a,b, 2017; Jacobs and Kinder, 2017, 2018). Recent examples for this approach are the prediction of the subjective beauty of single words (Jacobs, 2017), the literariness of metaphors (Jacobs and Kinder, 2018), or the most beautiful line of three Shakespeare sonnets (Jacobs, under revision[1]). Thus, using classifiers of the decision tree family (Geurts et al., 2006), Jacobs and Kinder identified a set of 11 features that could influence the literariness of metaphors, including their length, surprisal value, and sonority score (see below).

All these studies require training corpora as the basis for their computational predictions, and a particularly interesting challenge consists of finding or creating the optimal training corpus—especially for empirical scientific studies of *literature* (Jacobs, 2015c)—since standard corpora are not based on particularly literary texts. Recently, Bornet and Kaplan (2017) introduced a literary corpus of 35 French novels with over five million word tokens for a *named entity recognition* study, but in the fields of psycholinguistics and Neurocognitive Poetics, such specific corpora still are practically absent. An exception is the Shakespeare corpus (Shakespeare Online, http://www.shakespeare-online.com/sonnets/sonnetintroduction.html; cf. Jacobs et al., 2017) we recently used to compute the surprisal[2] values of entire sonnets, stanzas, or lines which are reliable and valid predictors of a number of response measures collected in empirical research on reading and literature, e.g., reading time or brain wave amplitudes (Frank, 2013). Surprisal computation requires a *language model*, usually based on trigrams (e.g., Jurafsky and Martin, 2000). However, it makes a big difference when trigram probabilities are computed on the basis of a nonliterary as compared to a literary or poetic training corpus, or when they are based on prose rather than poems (see below). As could be expected, when a contemporary corpus encompassing about six million sentences (SUBTLEX, Brysbaert and New, 2009) was used, a significantly higher mean surprisal (for all 154 sonnets) resulted than when the Shakespeare corpus was used (Jacobs et al., 2017). According to the Neurocognitive Poetics Model (Jacobs, 2011, 2015a,b), sonnets/lines/words with higher surprisal—and thus foregrounding potential—should more likely produce higher liking ratings, smaller eye movements, and longer fixation durations than sonnets low on surprisal. Data from a recent eye-tracking study using short literary stories support these predictions (van den Hoven et al., 2016). Regarding potential neuroimaging studies on sonnet reception, Jacobs et al. (2017) predicted a higher activation in several brain. Therefore areas, e.g., the left inferior temporal sulcus, bilateral superior temporal gyrus, right amygdala, bilateral anterior temporal poles, and right inferior frontal sulcus for sonnets with higher surprisal values. The choice of the training corpus and the language model is crucial for such predictions, the selection of the stimulus materials for empirical studies, and the evaluation of the theoretical model's descriptive accuracy and validity. A major goal of the Neurocognitive Poetics perspective is to develop and test training corpora of differing size, specificity, and representativeness in several languages (cf. Jacobs, under revision[1]).

In this paper, I describe a novel literary corpus assembled from the digitized books part of project Gutenberg (https://web.eecs.umich.edu/~lahiri/gutenberg_dataset.html), augmented by a Shakespeare corpus (Shakespeare Online, http://www.shakespeare-online.com/sonnets/sonnetintroduction.html; cf. Jacobs et al., 2017), henceforth called the Gutenberg Literary English Corpus (*GLEC*). The GLEC provides a collection of over 3,000 English texts from the Gutenberg project, spanning a wide range of genres, both fiction and non-fiction (novels, biographies, dramas, essays, short stories, novellas, tales, speeches and letters, science books, poetry; e.g., Austen, Bronte, Byron, Coleridge, Darwin, Dickens, Einstein, Eliot, Poe, Twain, Woolf, Wilde, Yeats) with about *12 million* sentences and *250 million* words.

MATERIALS AND METHODS: THE GLEC AND GEPC

The GLEC, i.e., the original Gutenberg texts augmented by the Shakespeare corpus, contains over 900 novels, over 500 short stories, over 300 tales and stories for children, about 200 poem collections, poems and ballads and about 100 plays, as well as over 500 pieces of non-fiction, e.g., articles, essays, lectures, letters, speeches, or (auto-)biographies. Except for the poetry collection subcorpus further explored in this paper and henceforth called the *Gutenberg English Poetry Corpus* (GEPC), these texts are not (yet) edited, shortened, or cleaned.

For the present analyses, I cleaned (in large part manually) all 116 texts making up the GEPC, e.g., by deleting duplicate poems,

[1] Jacobs, A.M. (2018). Computational stylistics and (Neuro-)cognitive poetics. *Scientific Study of Literature*. (under revision).

[2] Surprisal is the most common quantification of words' information content, usually computed as $\log_2[1/P(w_n \mid \text{context})]$ (Frank, 2013).

prefaces, introductions, content tables, and indices of first lines, postscripts, biographical, and author notes, as well as footnotes[3] or line and page numbers, and by separating poems from plays or essays (e.g., in Yeats texts), so that only the poems themselves remain in the texts without any piece of prose. This was important to obtain a valid "poetry-only" subcorpus and a valid *poetic language model* for comparison with poetic texts or text fragments, such as metaphors (Jacobs and Kinder, 2017, 2018). Without such cleaning, the computation of any *ngram model*, for instance, would be distorted by the prose parts. For the same reason, I also deleted poems in other languages than English, e.g., Lord Byrons "Sonetto di Vitorelli," PB Shelleys "Buona Notte," or TS Eliots "Dans le Restaurant."

In a second step, I concatenated all poetic texts written by a specific author which yielded a collection of 47 compound texts by the following authors: Aldous Huxley, Alexander Pope, Ambrose Bierce, Andrew Lang, Bret Harte, Charles Dickens, Charles Kingsley, DH Lawrence, Edgar Allan Poe, Elizabeth Barrett Browning, Ezra Pound, GK Chesterton, George Eliot, Herman Melville, James Joyce, James Russell Lowell, John Dryden, John Keats, John Milton, Jonathan Swift, Leigh Hunt, Lewis Carroll, Lord Byron, Lord Tennyson, Louisa May Alcott, Oscar Wilde, PB Shelley, Ralph Waldo Emerson, Robert Browning, Robert Frost, Robert Louis Stevenson, Robert Southey, Rudyard Kipling, Samuel Taylor Coleridge, Shakespeare, Sir Arthur Conan Doyle, Sir Walter Scott, Sir William Schwenck Gilbert, TS Eliot, Thomas Hardy, Walt Whitman, Walter de la Mare, William Blake, William Butler Yeats, William Dean Howells, William Makepeace Thackeray, and William Wordsworth. These 47 compound texts differ in a variety of surface and deep structure features, some of which are analyzed in the following sections. As can be seen in **Table A1** in the Appendix, text length also varies considerably across authors (exponential distribution with a median of 23,000 words): the top three authors are Lord Byron (~210,000 words), PB Shelley (~165,000), and Wordsworth (~115,000); the "flop" three are Alcott (<400), Pound (~1,200), and Joyce (~1,200). The majority of texts have less than 40,000 words. The entire GEPC comprises *1,808,160* words (tokens) and 41,857 types.

RESULTS: SOME EXEMPLARY ANALYSES OF THE GEPC

Like other fields (e.g., Computational Linguistics or Digital Humanities), Neurocognitive Poetics (Jacobs, 2015a) uses text corpora for many purposes, e.g., the abovementioned computation of surprisal values.

Other purposes are *similarity analyses*, which can be based on features extracted by latent semantic, topic, or sentiment analyses (e.g., Deerwester et al., 1990; Turney and Littman, 2003; Schmidtke et al., 2014a; Jacobs et al., 2015; Roe et al., 2016). Such features can then be used to train classifiers for identifying authors, periods of origin or main motifs, as well as for predicting ratings and other response data of poetic texts (e.g., van Halteren et al., 2005; Stamatatos, 2009; Jacobs and Kinder, 2017, 2018; Jacobs et al., 2017).

Similarity Analyses

As an example for a similarity analysis, **Figure 1** shows a multidimensional scaling representation of the 47 texts of the GEPC based on latent semantic analysis (document-term-matrix/DTM analysis[4]). Not surprisingly, this analysis reveals, e.g., that the "Lake poets" (e.g., Coleridge, Woodsworth) cluster together, or that some poets like Pound or Joyce stand out from the rest. The latter finding is related to the fact that the GEPC is most representative for poetry from the nineteenth century, a limitation discussed below.

Figure 2 shows a heat map comparing the 20 most significant topics (as extracted by *Non-Negative Matrix Factorization/NMF*; Pedregosa et al., 2011) for the 47 texts (a list of the 20 most significant words per topic is given in the Appendix). The color code is proportional to the probability of a given topic, i.e., all 20 values per author add up to 1.

The data summarized in **Figure 2** and the topic list (Appendix) reveal, e.g., that the (statistically) most important topic for Shakespeare's sonnets is topic #16 represented by the following 20 keyword stems: "natur spirit hath everi truth right hope think doth back find much faith art free round whole set drop." By contrast, topic #2 appears to be the most important for Lord Tennyson (keyword stems: "king knight arthur round queen answer mine lancelot saw lord mother arm call name thine hall among child hath speak").

Moreover, the data of **Figure 2** reveal that texts like those of Walt Whitman cover only four of the 20 topics (#1, 3, 4, and 17 have probabilities >0), whereas other authors such as Charles Dickens cover a large range of topics (i.e., 15/20 with $p > 0$). Such data can be used further in deeper analyses of generic poetic texts such as Shakespeare sonnets that look at topics important for esthetic success (e.g., Simonton, 1990) or for evoking specific affective and esthetic reader responses (Jacobs et al., 2017).

It should be noted[5] that at least a part of these topics represents authors and works as much as (or instead of) actual abstract concepts, likely because proper names were not filtered out. Thus, topic #18 (with terms like "cuchullain," "fintain," and "laigair") primarily describes Yeats's work and topic #2 Tennyson's "Idylls of the King." Comparative topic analyses using different algorithms and filters for, e.g., higher-frequency function words like "hath" or "without" can help determine the generality of such topics but are beyond the scope of this first paper introducing a new corpus. As argued elsewhere (see text footnote 1), the non-trivial interpretation of such data-driven learned topics can benefit from augmenting it by top-down conceptual tools such as the Cambridge Advanced Learner's Dictionary (Steyvers et al., 2011), expert knowledge in an iterative topic modeling process (Andrzejewski et al., 2009), or qualitative analyses concerning *thematic richness* or *symbolic imagery* (cf. Jacobs et al., 2017).

[3] For example, there were over 1,000 footnotes in "Lord Byrons Poetical Works Vol. 1" occupying a notable portion of the entire text.

[4] The DTM was based on sklearn-CountVectorizer with minimum term frequency = 1; maximum term frequency = 0.95, NLTK stopwords; stemmer = SnowballStemmer (cf. Bird et al., 2009; Pedregosa et al., 2011).

[5] I am indebted to one of the two reviewers for pointing this out.

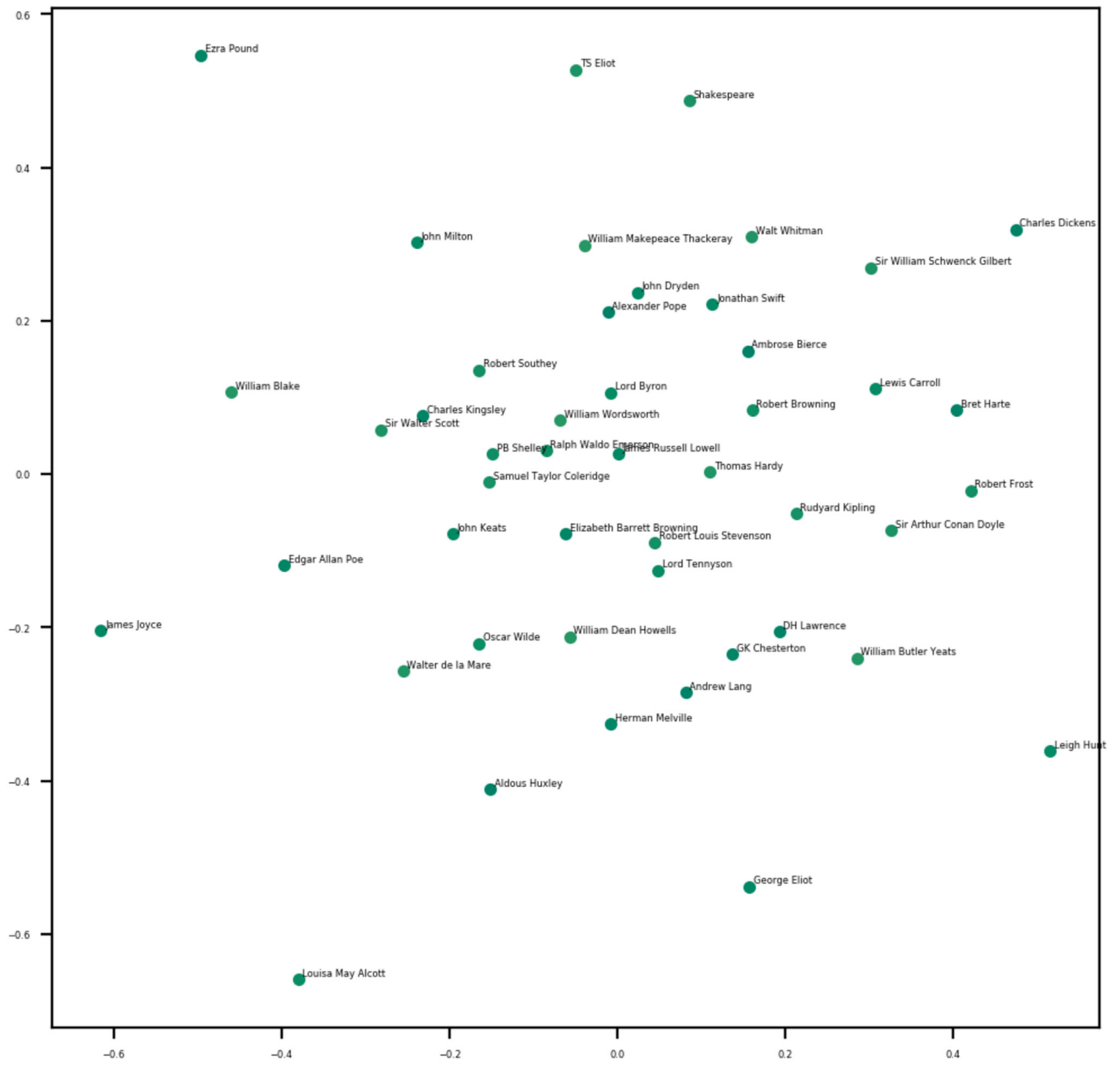

FIGURE 1 | Multidimensional scaling plot of text similarity for 47 authors from the Gutenberg English Poetry Corpus based on latent semantic analysis (document-term-matrix comparisons).

This can help in the creative task of finding a superordinate label for, say, the 20 words describing topic #16. As argued by Roe et al. (2016) who recently applied a topics analysis to the "French Encyclopedie," "the usefulness of a topic model does not necessarily rest on its ability to provide meaningful topics (a subjective categorization) for the corpus being analyzed, but rather on the multiplicity of perspectives it can generate and, as a result, on the potential for discovery that some of these topics can offer." In Neurocognitive Poetics, unsupervised topic modeling can also fulfill the role of a naïve "null-model" against which expert interpretations concerning focus and diversity (e.g., Vendler, 1997) can be gauged.

Comparing Word Uniqueness and Distinctiveness for Two Texts

A third and last example for how to use the present GEPC concerns a more detailed comparative analysis for a subset of the 47 texts including surface and semantic features. This is done for two authors with shorter texts of comparable length: Blake vs. Dickens (4,439 words vs. 3,758). We have recently

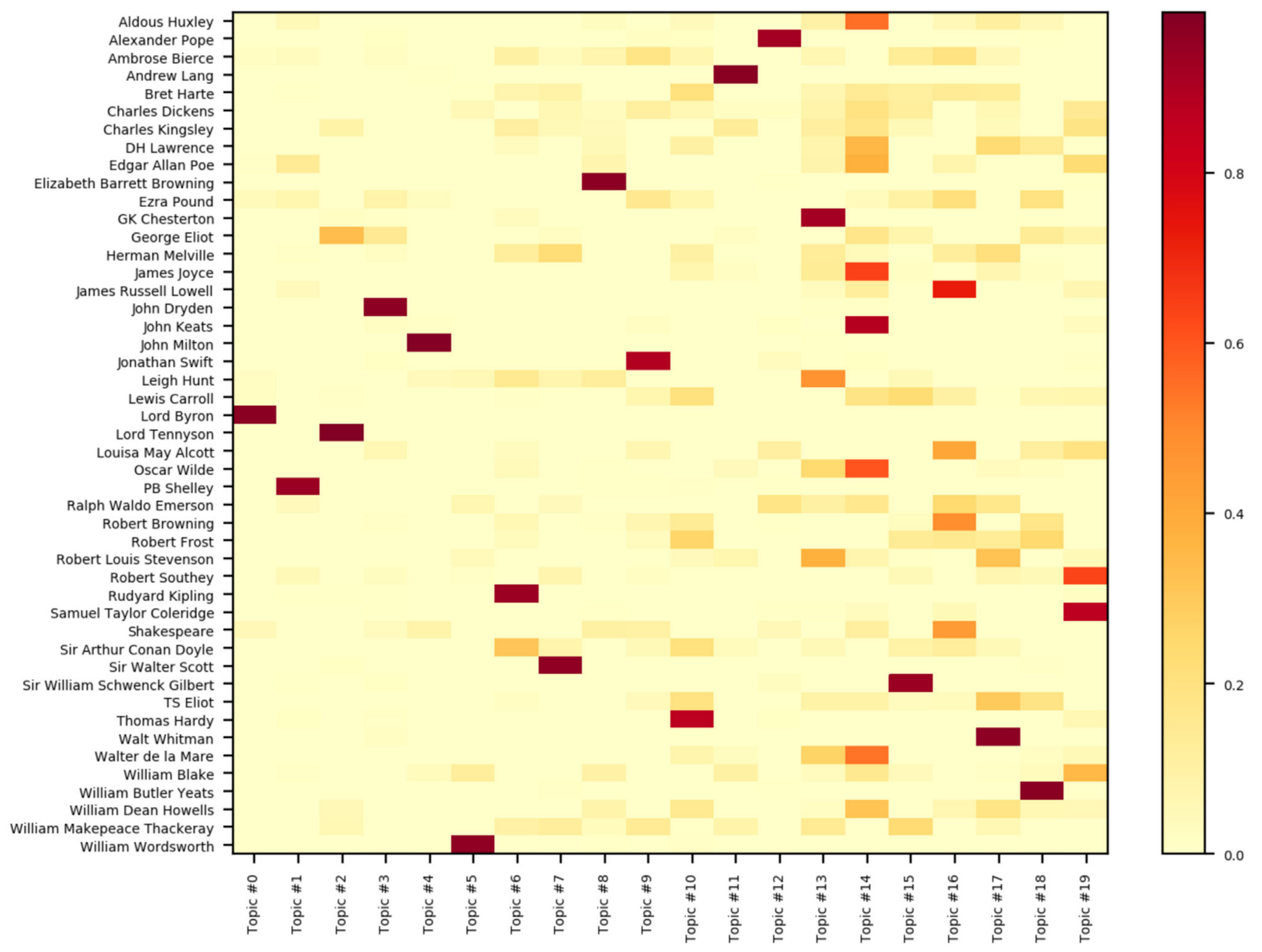

FIGURE 2 | Heat map of 20 main topics for the 47 texts of the Gutenberg English Poetry Corpus based on Non-Negative Matrix Factorization.

provided an extensive comparative QNA of all 154 Shakespeare sonnets looking at both surface and deep semantic features. For example, we compared features such as poem or line *surprisal*, *syntactic simplicity*, *deep cohesion*, or *emotion and mood potential* (Jacobs et al., 2017). As an example for another interesting feature not considered in our previous study, here I will focus on word distinctiveness or *keyness*. In computing this feature, I closely followed the procedure proposed in DARIAH—Digital Research Infrastructure for the Arts and Humanities; https://de.dariah.eu/tatom/feature_selection.html. According to DARIAH's operationalization, one way to consider words as distinctive is when they are found exclusively in texts associated with a single author (or group). For example, if Dickens uses the word "squire" in the present GEPC and Blake never does, one can count "squire" as distinctive or unique (in this comparative context). *Vice versa*, the word "mother" is distinctive in this GEPC comparison, because Dickens never uses it (see **Table 1**).

Identifying unique words simply requires to calculate the average rate of word use across all texts for each author and then to look for cases where the average rate is zero for one author. Based on the DTMs for both texts, this yielded the following results.

TABLE 1 | Five unique words with usage rates (1/1,000) in Blake's and Dickens' poems of the GEPC.

Author/words	SQUIRE	LUCI	MOTHER	FINE	LAMB
Blake	0	0	7.9	0	7.2
Dickens	11.2	8.7	0	7.5	0

Another approach to measuring *keyness* is to compare the average rate at which authors use a word by calculating the difference between the rates. Using this measure, I calculated the top five distinctive words in the Blake–Dickens comparison by dividing the difference in both authors' average rates by the average rate across all 47 authors.

Thus, appearing only once in the entire text, Dickens' word stem "outgleam" in the line "Behold outgleaming on the angry main!" appears to be distinctive, much as the other four word stems in **Table 2**.

A final quantitative comparison inspired by DARIAH's approach to determining word distinctiveness uses a Bayesian group or an author comparison. It involves estimating the belief about the observed word frequencies to differ significantly by using a probability distribution called the sampling model. This assumes the rates to come from two different normal distributions, and the question to be answered is how confident one is that the means of the two normal distributions are different. The degree of confidence (i.e., a Bayesian probability), that the means are indeed different, then is another probabilistic measure of distinctiveness.

Using a *Gibbs sampler* to get a distribution of posterior values for δ[6] which is the variable estimating the belief about the difference in authors' word usage (for details, see https://de.dariah.eu/tatom/feature_selection.html., cf. Burrows, 2002), I computed the probability that using the words "squire" and "fine" (both more characteristic of Dickens' poems than of Blake's) are likely to be zero (see **Table 3**).

TABLE 2 | Top five distinctive words (stems) with usage rates (1/1,000) in Blake's and Dickens' poems of the GEPC.

Author/words	dol'	outgleam	chalon	toor	vithin
Blake	0	0	0	0	0
Dickens	0.83	0.41	0.41	0.83	0.41

TABLE 3 | Bayesian probability estimates (based on 2,000 samples) for two distinctive words (SQUIRE, FINE) with usage rates (1/1,000) in Blake's and Dickens' poems of the GEPC.

	SQUIRE	FINE
p ($\delta < 0$)	0.23	0.09
Blake average	0	0
Dickens average	11.2	7.5

According to this Bayesian analysis, "squire" appears more distinctive of Dickens' poetry than "fine," but since both words do not produce a high probability of differing from zero, I would not put much belief in them being specifically characteristic of Dickens in the GEPC (although they are most distinctive in comparison to Blake, see **Tables 1** and **3**). This Bayesian "feature selection" method can be extended to every word occurring in a corpus producing a useful ordering of characteristic words (for details, see https://de.dariah.eu/tatom/feature_selection.html).

Comparing Two Individual Poems

The above analyses dealt with the entire GEPC or two poem collections, respectively. Next, I focus on a more detailed—purely descriptive—comparison of two short individual texts from the GEPC that are far apart from each other (and the rest of the poems) in the similarity graph shown in **Figure 1**: George Eliot's poem "How Lisa Loved the King" and James Joyce's "Chamber Music." I will give just a few illustrative statistics both for surface and for deeper semantic features that are of potential use in Digital Humanities and Neurocognitive Poetics studies (for review on the latter, see Jacobs, 2015a; Jacobs et al., 2017).

Two features that are often used as indicators of linguistic complexity, poetic quality, or esthetic success are *lexical diversity*—measured by the *type–token ratio*—and *adjective–verb quotient*: for example, "better" Shakespeare sonnets are distinguished by a higher type-token ratio, more unique words, and a higher adjective–verb quotient (e.g., Simonton, 1989). The number of types can also be considered a coestimate of the size of an authors' (active) mental lexicon and vocabulary profile. As can be seen in columns 2 and 3 of **Table 4**, both poems descriptively do not differ much on these features.

Looking at the three most frequent nouns, verbs, and adjectives, as well as significant bi- and trigram collocations in columns 4 and 5, the keywords suggest that both poems have much to say about one of three favorite poetry motifs, i.e., *love*. This is also evident from the two lexical dispersion plots shown in **Figure 3,** which show, among others, that "love" appears well distributed across the entire poems, never letting the reader forget the poems' central motif.

TABLE 4 | Some exemplary statistics for two poems.

Author	Nbr. of word tokens/ types/hapaxes/type–token ratio (lexical diversity)	Nbr. of nouns, verbs, adjectives/adjective–verb quotient	Most freq. nouns, verbs, adjectives	Most freq. bi- and trigram collocations	Mean sonority score	Mean positive and negative valence, and arousal/most positive, negative, and arousing word
Eliot	2,702, 1,467, 1,014, 0.5	1,111, 686, 642, 0.93	LOVE (19), LIFE (15), SOUL (12), love (7), see (5), live (3) little (13), high (9), good (9)	"King Pedro" (4), "day might " (2), "death tell" (2), "Six hundred years " (2), "Hundred years ago" (2), "T gentle Lisa" (1)	5.19	1.01, 0.84, 2.01 happiness, shame, happiness
Joyce	1,221, 654, 447, 0.53	507, 313, 270, 0.86	LOVE (23), HEART (18), AIR (9) love (7), come (3), sleep (2) sweet (13), soft (9), fair (5)	"true love" (4), "long hair" (3), "pretty air" (3), "combing long hair" (2), "would sweet bosom" (2), "singing merry air" (2)	5.26	1.02, 0.85, 2.03 happiness, sadness, happiness

[6] It represents half the difference between the population means for the distributions characterizing word rates in Blake and Dickens.

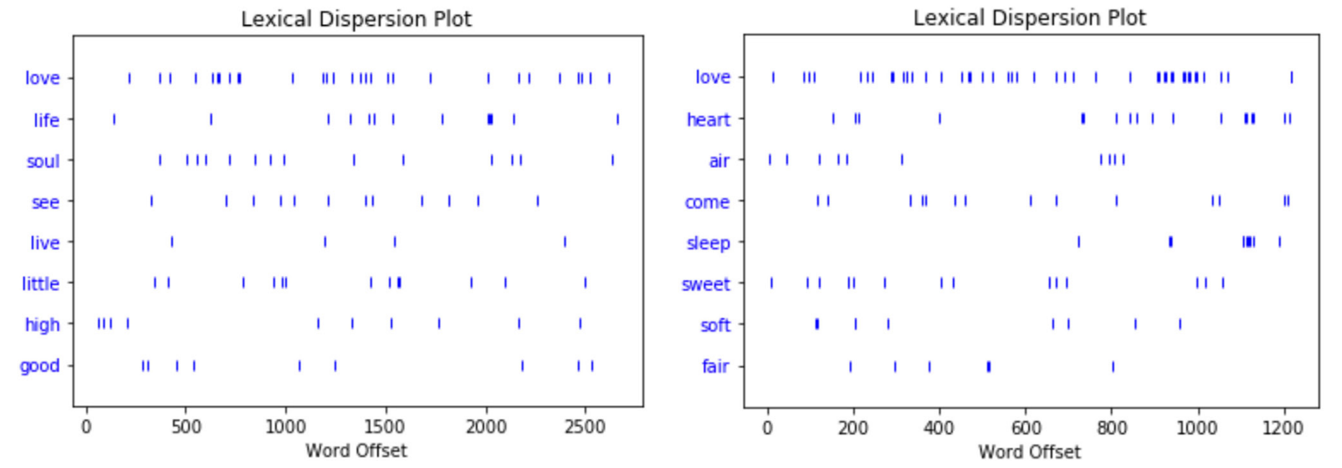

FIGURE 3 | Lexical dispersion plots for the three most frequent nouns, verbs, and adjectives in George Eliot's poem "How Lisa Loved the King" (right panel) and James Joyce's "Chamber Music" (left panel). Note that the word "love" is counted as both noun and verb here.

Poetic language expertly plays with the sound-meaning nexus, and our group has provided empirical evidence that sublexical phonological features play a role in (written) poetry reception (Schrott and Jacobs, 2011; Aryani et al., 2013, 2016; Schmidtke et al., 2014b; Jacobs, 2015b,c; Jacobs et al., 2015, 2016b; Ullrich et al., 2017). A sublexical phonological feature with poetic potential is the *sonority score* (Jacobs, 2017; Jacobs and Kinder, 2018; see Appendix A for details). It is based on the notion of sonority profile (cf. Clements, 1990; Stenneken et al., 2005) which rises maximally toward the peak and falls minimally toward the end, proceeding from left to right, for the universally preferred syllable type (Clements, 1990, p. 301). Through a process of more or less unconscious *phonological recoding,* text sonority may play a role even in silent reading (Ziegler and Jacobs, 1995; Braun et al., 2009) and especially in reading poetic texts (Kraxenberger, 2017). Column six of **Table 4** shows that the two poems differ little in their global sonority score. At a finer-grained level of individual lines or stanzas, sonority could still notably differ, however, and implicitly affect readers' affective-esthetic evaluation (cf. Jacobs and Kinder, 2018).

An important task for QNA-based Neurocognitive Poetics studies is *sentiment analysis*, i.e., to estimate the emotional valence or mood potential of verbal materials (e.g., Jacobs et al., 2017). In principle, this is done with either of two methodological approaches: using word lists that provide values of word valence or arousal based on human rating data (e.g., Jacobs et al., 2015), or applying a method proposed by Turney and Littman (2003) based on associations of a target word with a set of *labels*, i.e., keywords assumed to be prototypical for a certain affect or emotion. Following previous research (Westbury et al., 2014), I computed the lexical features *valence* and *arousal* according to a procedure described in Appendix B.

The mean values in the rightmost column of **Table 4** indicate that at this *global level,* both poems practically do not differ on any of these three affective features. This can be visualized for the entire poems by the 3D plots of the principal components extracted from the three variables for all words in the poems: descriptively, they appear very similar (see **Figure 4**). All other things being equal, this suggests that, e.g., human ratings of the global affective meaning of both poems should not differ significantly (cf. Aryani et al., 2016). Of course, at the *local level,* a deeper qualitative analysis of both poems may reveal that they do in fact inhabit completely different esthetic universes which influence such ratings', as pointed out by one reviewer who also noted that "the formal differences (Joyce's variety of line lengths, meters, and stanza shapes vs. Eliot's fairly straight ahead iambic pentameter) have a strong impact on the atmosphere the poems create." As we have repeatedly argued elsewhere (Jacobs, 2015b; Jacobs et al., 2017; Abramo et al., under revision[7]), QNA-based text analyses like these global affect scores should be complemented by qualitative analyses of style figures—done by interdisciplinary experts—at all levels of the 4 × 4 matrix proposed in Jacobs (2015b), i.e., metric, phonological, morpho-syntactic, semantic, as well as sublexical, lexical, interlexical, and supralexical. The *Foregrounding Assessment Matrix* recently proposed by Abramo et al. (under revision)[7] is such a useful tool that allows to identify *density fields* of overlapping style figures at several levels, e.g., sublexical–phonological (alliteration) and interlexical–semantic (metaphor). As a promising first result, the combined qualitative–quantitative analysis of Shakespeare's sonnet 60 allowed these authors to predict the keyword score (i.e., words marked by readers as being "keywords" for understanding the sonnet) with an accuracy of about 90%.

It is thus important to note that QNA-based analyses like those in **Table 4** are not meant to replace deep qualitative analyses of texts like Vendler's (Vendler, 1997) interpretation of Shakespeare's sonnets. However, for designing Neurocognitive Poetics studies,

[7] Abramo, F., Gambino, R., Pulvirenti, G., Xue, S., Sylvester, T., Mangen, A., et al. (2018). A Qualitative–Quantitative Analysis of Shakespeare's Sonnet 60. Style (under revision).

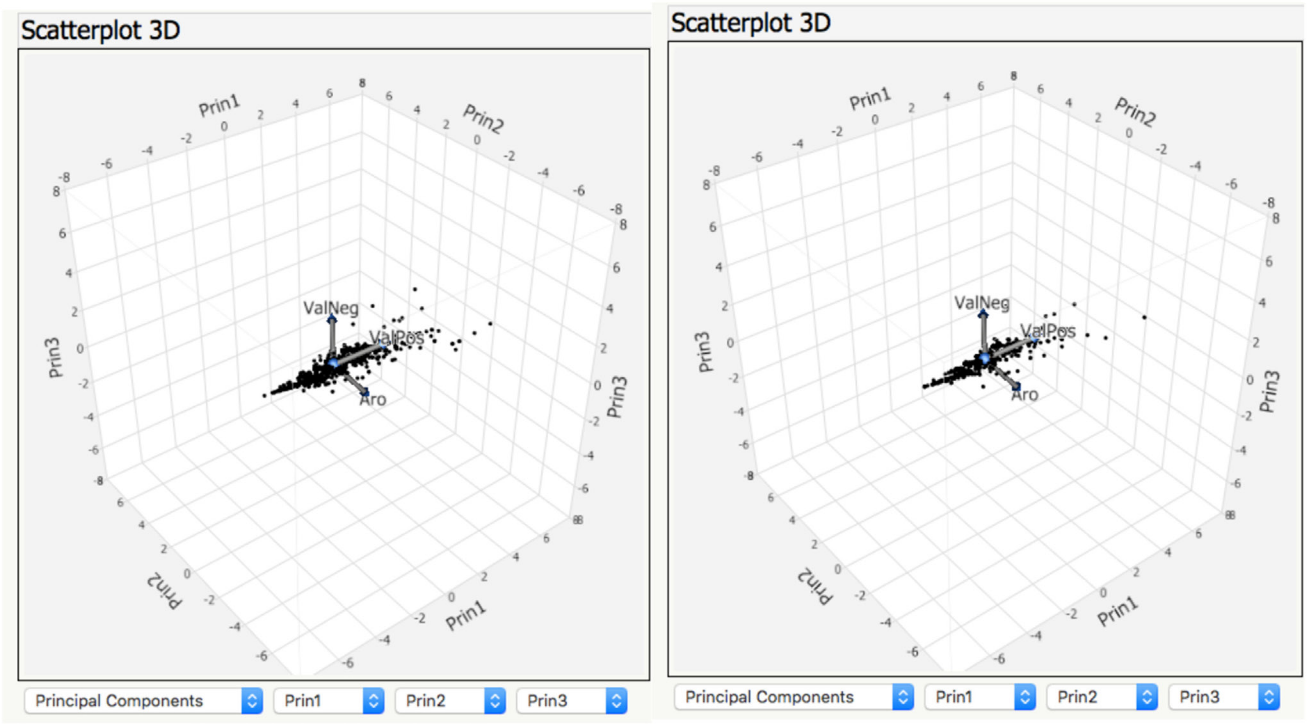

FIGURE 4 | 3D scatterplot for the principal components extracted from the valence and arousal values of the words in George Eliot's poem "How Lisa Loved the King" (left panel) and James Joyce's "Chamber Music" (right panel).

which involve selecting and matching complex verbal materials on a variety of feature dimensions, they are a necessity (cf. Jacobs and Willems, 2018).

DISCUSSION

In this paper, I have briefly described a relatively big corpus of English literary texts, the GLEC, for use in studies of Computational Linguistics, Digital Humanities, or Neurocognitive Poetics. As a whole, the GLEC requires further processing (e.g., cleaning, regrouping according to subgenres, etc.) before it can be used as a training and/or test corpus for future studies. Using a smaller subcorpus already cleaned and consisting of 116 poetry collections, poems, and ballads from 47 authors, i.e., the GEPC, I presented a few exemplary QNA studies in detail. In these explorations of the GEPC, I showed how to use similarity and topic analyses for comparing and grouping texts, several methods for identifying distinctive words, and procedures for quantifying important features that can influence reader responses to literary texts, e.g., lexical diversity, sonority score, valence, or arousal. The GEPC thus can be applied to a variety of research questions such as authorship and period of origin classifications (cf. Stamatatos, 2009), the prediction of beauty ratings for metaphors (e.g., Jacobs and Kinder, 2017, 2018), or the design of neuroimaging studies using literary stimuli (e.g., Bohrn et al., 2013; O'Sullivan et al., 2015).

The still relatively rare application of corpus-based QNA to poetry is an integral part of the Neurocognitive Poetics Perspective (e.g., Jacobs, 2015b,c; O'Sullivan et al., 2015; Willems and Jacobs, 2016; Jacobs and Willems, 2018; Jacobs et al., 2017; Nicklas and Jacobs, 2017; Jacobs and Kinder, 2018), because it offers the possibility of neurocognitive experiments with complex, natural verbal stimuli that can vary on a plethora of features (e.g., >70 in Jacobs & Kinder's recent metaphor study). While being a first necessary step for state-of-the art statistical data analyses (e.g., in eye tracking or fMRI studies), augmenting QNA with interdisciplinary expert qualitative text analyses (e.g., see text footnote 7) is also necessary, because the rich esthetics of poetry is based on a complex author-(con-)text-reader nexus QNA tools alone cannot cope with. While the detection, dynamic development (across poem parts), and interpretation of metaphors are a case in point, the above data-driven topic analyses also indicate both the potential and limitations of QNA not only for poetry but also for prose or scientific texts, where they can be considered a useful complement to traditional methods of close reading (e.g., Roe et al., 2016).

The GLEC and GEPC are two of many available training corpora and can be compared or also combined with much larger general corpora, such as *ukwac* (Baroni et al., 2009). I have discussed strengths and limits of a dozen training corpora useful for Neurocognitive Poetics and computational stylistics elsewhere (see text footnote 1). The obvious limitation of the present corpus lies in its texts being relatively "old": due to copyright issues, the GLEC and GEPC contain only texts from 1623 to 1952, the majority of the GEPC stemming from the nineteenth century

(Median = 1885). This limitation can at least partly be overcome by merging the GEPC with the contemporary ukwac corpus (>2 billion words), for example. To what extent this *ukwac-GEPC* merger is appropriate for studies using more modern or contemporary prose and poetry texts is an open theoretical and empirical question to be addressed in future comparative research. The successful application of the GLEC as a reliable language model (with a hit rate of 100%) for the computation of the surprisal values of 464 metaphors which also included contemporary ones (Katz et al., 1988) is encouraging in this respect (Jacobs and Kinder, 2018).

The development of appropriate open-access training corpora—which are both sufficiently specific and representative for the research materials and reader population under investigation—is one of four general desiderata of current computational stylistics and Neurocognitive Poetics (see text footnote 1) together with the development of combined qualitative–quantitative narrative analysis (Q2NA) and machine-learning tools for feature extraction, standard ecologically valid literary test materials (Hanauer, 2017), and open-access reader response data banks. These developments will not replace the art of close reading and interpreting literary texts, but—paraphrasing Hanauer—they may well lead to "stronger and more generalizable hypotheses about literary phenomena in the future" and thus attract and generate more cross-disciplinary research which ideally leads to a cross-fertilization between the humanities and sciences in the domain of literature, much in the spirit of Zipf (1932), Turner and Poeppel (1983), or Michel et al. (2011).

AUTHOR CONTRIBUTIONS

AJ conceived and wrote the MS and carried out all original work (data collection and analyzing, python programming, etc.) reported herein.

REFERENCES

Andrzejewski, D., Zhu, X., and Craven, M. (2009). Incorporating domain knowledge into topic modeling *via* Dirichlet forest priors. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, 25–32. New York, NY: ACM.

Aryani, A., Jacobs, A.M., and Conrad, M. (2013). Extracting salient sublexical units from written texts: "Emophon," a corpus-based approach to phonological iconicity. *Frontiers in Psychology* 4:654. doi:10.3389/fpsyg.2013.00654

Aryani, A., Kraxenberger, M., Ullrich, S., Jacobs, A.M., and Conrad, M. (2016). Measuring the basic a ective tone of poems *via* phonological saliency and iconicity. *Psychology of Aesthetics, Creativity, and the Arts* 10: 191–204. doi:10.1037/aca0000033

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky Wide Web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43: 209–26. doi:10.1007/s10579-009-9081-4

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media, Inc.

Bohrn, I.C., Altmann, U., Lubrich, O., Menninghaus, W., and Jacobs, A.M. (2013). When we like what we know—a parametric fMRI analysis of beauty and familiarity. *Brain and Language* 124: 1–8. doi:10.1016/j.bandl.2012.10.003

Bornet, C., and Kaplan, F. (2017). A simple set of rules for characters and place recognition in French novels. *Frontiers in Digital Humanities* 4:6. doi:10.3389/fdigh.2017.00006

Braun, M., Hutzler, F., Ziegler, J.C., Dambacher, M., and Jacobs, A.M. (2009). Pseudo homophone effects provide evidence of early lexico-phonological processing in visual word recognition. *Human Brain Mapping* 30: 1977–89. doi:10.1002/hbm.20643

Brysbaert, M., and New, B. (2009). Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41: 977–90. doi:10.3758/BRM.41.4.977

Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* 17: 267–87. doi:10.1093/llc/17.3.267

Clements, G.N. (1990). The role of sonority in core syllabification. In *Papers in Laboratory Phonology I. Between the Grammar and Physics of Speech*, Edited by J. Kingston and M.E. Beckman, 283–333. Cambridge: CUP.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41: 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Frank, S.L. (2013). Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science* 5: 475–94. doi:10.1111/tops.12025

Ganascia, J.-G. (2015). The logic of the big data turn in digital literary studies. *Frontiers in Digital Humanities* 2:7. doi:10.3389/fdigh.2015.00007

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning* 63: 3–42. doi:10.1007/s10994-006-6226-1

Hanauer, D. (2017). Towards a critical mass of accumulated knowledge in the field of scientific literary studies. *Scientific Study of Literature* 7: 1–3. doi:10.1075/ssol.7.1.01edi

Jacobs, A.M. (2011). Neurokognitive Poetik: Elemente eines Modells des literarischen Lesens [Neurocognitive poetics: Elements of a model of literary reading]. In *Gehirn und Gedicht: Wie wir unsere Wirklichkeiten konstruieren [Brain and Poetry: How We Construct Our Realities]*, Edited by R. Schrott and A.M. Jacobs, 492–520. Munich: Carl Hanser.

Jacobs, A.M. (2015a). Neurocognitive poetics: methods and models for investigating the neuronal and cognitive–affective bases of literature reception. *Frontiers Human Neuroscience* 9:186. doi:10.3389/fnhum.2015.00186

Jacobs, A.M. (2015b). Towards a neurocognitive poetics model of literary reading. In *Cognitive Neuroscience of Natural Language Use*, Edited by R. Willems, 135–159. Cambridge, England: Cambridge University Press.

Jacobs, A.M. (2015c). The scientific study of literary experience: sampling the state of the art. *Scientific Study of Literature* 5: 139–70. doi:10.1075/ssol.5.2.01jac

Jacobs, A.M. (2017). Quantifying the beauty of words: a neurocognitive poetics perspective. *Frontiers in Human Neuroscience* 11:622. doi:10.3389/fnhum.2017.00622

Jacobs, A.M., Hofmann, M.J., and Kinder, A. (2016a). On elementary affective decisions: to like or not to like, that is the question. *Frontiers Psychology* 7:1836. doi:10.3389/fpsyg.2016.01836

Jacobs, A.M., Lüdtke, J., Aryani, A., Meyer-Sickendiek, B., and Conrad, M. (2016b). Mood- empathic and aesthetic responses in poetry reception: a model-guided, multilevel, multimethod approach. *Scientific Study of Literature* 6: 87–130. doi:10.1075/ssol.6.1.06jac

Jacobs, A.M., and Kinder, A. (2017). The brain is the prisoner of thought: a machine-learning assisted quantitative narrative analysis of literary metaphors for use in Neurocognitive Poetics. *Metaphor and Symbol* 32: 139–60. doi:10.1080/10926488.2017.1338015

Jacobs, A.M., and Kinder, A. (2018). What makes a metaphor literary? Answers from two computational studies. *Metaphor and Symbol*. in press.

Jacobs, A.M., Schuster, S., Xue, S., and Lüdtke, J. (2017). What's in the brain that ink may character ….: a quantitative narrative analysis of Shakespeare's 154 sonnets for use in neurocognitive poetics. *Scientific Study of Literature* 7: 4–51. doi:10.1075/ssol.7.1.02jac

Jacobs, A.M., Võ, M.L.-H., Briesemeister, B.B., Conrad, M., Hofmann, M.J., Kuchinke, L., et al. (2015). 10 years of BAWLing into affective and aesthetic processes in reading: what are the echoes? *Frontiers in Psychology* 6: 714. doi:10.3389/fpsyg.2015.00714

Jacobs, A.M., and Willems, R.M. (2018). The fictive brain: neurocognitive correlates of engagement in literature. *Review of General Psychology*. in press. doi:10.1037/gpr0000106

Jakobson, R., and Lévi-Strauss, C. (1962). “Les chats” de Charles Baudelaire. *L'homme* 2: 5–21. doi:10.3406/hom.1962.366446

Jurafsky, D., and Martin, J.H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall.

Katz, A., Paivio, A., Marschark, M., and Clark, J. (1988). Norms for 204 literary and 260 non-literary metaphors on psychological dimensions. *Metaphor and Symbolic Activity* 3: 191–214. doi:10.1207/s15327868ms0304_1

Kraxenberger, M. (2017). *On Sound-Emotion Associations in Poetry*. Ph.D. thesis, Freie University, Berlin.

Leech, G.N. (1969). *A Linguistic Guide to English Poetry*. London, UK: Longman.

Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K, Google Books Team, et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science* 331: 176–82.

Mitchell, T.M. (1997). *Machine Learning*. New York: McGraw-Hill.

Moretti, F. (2005). *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso.

Nicklas, P., and Jacobs, A.M. (2017). Rhetorics, neurocognitive poetics and the aesthetics of adaptation. *Poetics Today* 38: 393–412. doi:10.1215/03335372-3869311

O'Sullivan, N., Davis, P., Billington, J., Gonzalez-Diaz, V., and Corcoran, R. (2015). “Shall I compare thee”: the neural basis of literary awareness, and its benefits to cognition. *Cortex; A Journal Devoted to the Study of the Nervous System and Behavior* 73: 144–57. doi:10.1016/j.cortex.2015.08.014

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). scikit-learn: machine learning in Python. *The Journal of Machine Learning Research* 12: 2825–30.

Roe, G., Gladstone, C., and Morrissey, R. (2016). Discourses and disciplines in the enlightenment: topic modeling the french encyclopeédie. *Frontiers of Digital Humanities* 2: 8. doi:10.3389/fdigh.2015.00008

Schmidtke, D.S., Schröder, T., Jacobs, A.M., and Conrad, M. (2014a). ANGST: affective norms for German sentiment terms, derived from the affective norms for English words. *Behavior Research Methods* 46: 1108–18. doi:10.3758/s13428-013-0426-y

Schmidtke, D.S., Conrad, M., and Jacobs, A.M. (2014b). Phonological iconicity. *Frontiers in Psychology* 5:80. doi:10.3389/fpsyg.2014.00080

Schrott, R., and Jacobs, A.M. (2011). *Gehirn und Gedicht: Wie wir unsere Wirklichkeiten konstruieren (Brain and Poetry: How We Construct Our Realities)*. München, Germany: Hanser.

Simonton, D.K. (1989). Shakespeare's Sonnets: a case of and for single–case historiometry. *Journal of Personality* 57: 695–721. doi:10.1111/j.1467-6494.1989.tb00568.x

Simonton, D.K. (1990). Lexical choices and aesthetic success: a computer content analysis of 154 Shakespeare sonnets. *Computers and the Humanities* 24: 254–64.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science & Technology* 60: 538–56. doi:10.1002/asi.21001

Stenneken, P., Bastiaanse, R., Huber, W., and Jacobs, A.M. (2005). Syllable structure and sonority in language inventory and aphasic neologisms. *Brain & Language* 95: 280–92. doi:10.1016/j.bandl.2005.01.013

Steyvers, M., Smyth, P., and Chemuduganta, C. (2011). Combining background knowledge and learned topics. *Topics in Cognitive Science* 3: 18–47. doi:10.1111/j.1756-8765.2010.01097.x

Stockwell, P. (2002). *Cognitive Poetics: An Introduction*. London: Routledge.

Tsur, R. (1983). *What is Cognitive Poetics?* Tel aviv: Katz Research Institute for Hebrew Literature.

Turner, F., and Poeppel, E. (1983). The neural lyre: poetic meter, the brain and time. *Poetry Magazine* 12: 277–309.

Turney, P.D., and Littman, M.L. (2003). Measuring praise and criticism: inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)* 21: 315–46. doi:10.1145/944012.944013

Ullrich, S., Aryani, A., Kraxenberger, M., Jacobs, A.M., and Conrad, M. (2017). On the relation between the general affective meaning and the basic sublexical, lexical, and interlexical features of poetic texts—a case study using 57 poems of H. M. Enzensberger. *Frontiers in Psychology* 7:2073. doi:10.3389/fpsyg.2016.02073

van den Hoven, E., Hartung, F., Burke, M., and Willems, R. (2016). Individual differences in sensitivity to style during literary reading: insights from eye-tracking. *Collabra: Psychology* 2: 1–16. doi:10.1525/collabra.39

van Halteren, H., Baayen, R.H., Tweedie, F., Haverkort, M., and Neijt, A. (2005). New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics* 12: 65–77. doi:10.1080/09296170500055350

Vendler, H. (1997). *The Art of Shakespeare's Sonnets*. Cambridge, MA: Harvard University Press.

Westbury, C., Keith, J., Briesemeister, B.B., Hofmann, M.J., and Jacobs, A.M. (2014). Avoid violence, rioting, and outrage; approach celebration, delight, and strength: using large text corpora to compute valence, arousal, and the basic emotions. *Quarterly Journal of Experimental Psychology* 68: 1599–622. doi:10.1080/17470218.2014.970204

Willems, R., and Jacobs, A.M. (2016). Caring about Dostoyevsky: the untapped potential of studying literature. *Trends in Cognitive Sciences* 20: 243–5. doi:10.1016/j.tics.2015.12.009

Ziegler, J.C., and Jacobs, A.M. (1995). Phonological information provides early sources of constraint in the processing of letter strings. *Journal of Memory and Language* 34: 567–93. doi:10.1006/jmla.1995.1026

Ziegler, J.C., Stone, G.O., and Jacobs, A.M. (1997). What is the pronunciation for -ough and the spelling for/u/? A database for computing feedforward and feedback consistency in English. *Behavior Research Methods, Instruments, and Computers* 29: 600–18. doi:10.3758/BF03210615

Zipf, G.K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA: Harvard University Press.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer, CC, and handling Editor declared their shared affiliation.

APPENDIX

Top 20 Topics with 20 Keyword Stems

Topic 0: much hath name without juan vain hope thine everi less spirit mine doge blood call form youth art think sinc

Topic 1: spirit power hope cloud human deep mountain wave thine among art beneath star blood natur everi wild form around breath

Topic 2: king knight arthur round queen answer mine lancelot saw lord mother arm call name thine hall among child hath speak

Topic 3: king everi wit power much law art kind fate name princ age find grace foe natur arm show without sinc

Topic 4: thir son hath power hast angel self spirit lord stood much hell art find arm without glori less king father

Topic 5: natur marmaduk hath hope power among hill joy oswald everi side round spirit name child sight wood seen father human

Topic 6: back lord king call mother son stand done work run round follow hold soldier watch fight get road tell mine

Topic 7: marmion lord wild king dougla deep vain war saint name band roderick mountain arm knight foe show blood hill everi

Topic 8: angel spirit back mine child thine slowli think toll speak name stand curs adam round sin breath call strong lip

Topic 9: everi wit find think show call much sinc tell lord pleas name muse without better poet virtu art court learn

Topic 10: call near back think mine stood show saw stand yes ere none much sinc hous yea wait name tree woman

Topic 11: king lord bonni ballad son helen green fell three john child father saw queen set gray tell war back round

Topic 12: wit natur everi pleas art fool virtu name learn sens prais vain muse grace fame pride reason call poet lord

Topic 13: king star tree lord sword hous grow gray stand green saw bird hill stood alfr break deer stone wild fell

Topic 14: green breath lip round everi pain silver kiss gentl tell thine feet deep tree doth wide star spirit wild blue

Topic 15: ballad everi kind think tell peter boy name maid pretti marri captain maiden get much doubt willow call tri plan

Topic 16: natur spirit hath everi truth right hope think doth back find much faith art free round whole set drop poet

Topic 17: citi everi hous think noth bodi river state ship stand back shore other wood women pioneer arm star wait mother

Topic 18: king pupil fool cuchullain find hous fintain noth woman call conal run prais barach put anoth laegair tell bodi think

Topic 19: hope joy mother spirit beneath round maid child name power wild deep form breast father natur cloud youth lord gaze

A. Computing the Sonority Score

Following previous work (Stenneken et al., 2005; Jacobs and Kinder, 2018) and considering that here we deal with written instead of spoken words, I used a simplified index based on the sonority hierarchy of English phonemes which yields 10 ranks: [a] > [e o] > [i u j w] > [ɾ] > [l] > [m n ŋ] > [z v] > [f θ s] > [b d g] > [p t k]. Each

TABLE A1 | List of authors with example texts in the Gutenberg Literary English Corpus (GLEC) and total text lengths in the Gutenberg English Poetry Corpus (GEPC).

Author	Nbr. of texts	Example text in GLEC, year of publication	*GEPC text, length (nbr. of words)*
1. Abraham Lincoln	16	Lincoln's First Inaugural Address, 1861	–
2. Agatha Christie	2	The Secret Adversary, 1922	–
3. Albert Einstein	2	Relativity/The Special and General Theory, 1916	–
4. Aldous Huxley	3	Crome Yellow, 1921	*The Defeat of Youth and Other Poems, 4,616*
5. Alexander Pope	3	The Rape of the Lock and Other Poems, 1875	*The Poetical Works, 82,870*
6. Alfred Russel Wallace	5	Is Mars Habitable? 1907	–
7. Ambrose Bierce	18	A Cynic Looks at Life, 1912	*Black Beetles in Amber, 23,815*
8. Andrew Lang	60	Historical Mysteries, 1904	*A Collection of Poems, 46,466*
9. Anthony Trollope	71	The Eustace Diamonds, 1871	–
10. Arnold J. Toynbee	1	Turkey/A Past and a Future, 1917	–
11. Baronness Orczy	16	The Tangled Skein, 1907	–
12. Beatrix Potter	1	A Collection of Beatrix Potter Stories, 1902	–
13. Benjamin Disraeli	17	Vivian Grey, 1826	–
14. Benjamin Franklin	4	Autobiography of Benjamin Franklin, Version 4, 1791	–
15. Bertrand Russell	8	The Analysis of Mind, 1921	–
16. Bram Stoker	6	Dracula, 1897	–
17. Bret Harte	58	The Queen of the Pirate Isle, 1886	*East and West, 6,737*
18. Charles Darwin	20	The Expression of Emotion in Man and Animals, 1859	–
19. Charles Dickens	60	Oliver Twist, 1837	*The Poems and Verses, 3,758*
20. Charles Kingsley	44	True Words for Brave Men, 1884	*Poems, 14,391*
21. Charlotte Bronte	4	Jane Eyre, 1847	–
22. DH Lawrence	19	Women in Love, 1920	*Collected Poems, 19,820*
23. Edgar Allen Poe	11	The Masque of the Red Death, 1842	*Complete Poetical Works, 8,117*
24. Edgar Rice Burroughs	25	Tarzan of the Apes, 1912	–
25. Edmund Burke	15	Burke's Speech on Conciliation with America, 1775	–
26. Edward P Oppenheim	53	The Zeppelin's Passenger, 1918	–
27. Elizabeth B Browning		Sonnets From the Portuguese, 1850	*The Poetical Works, 59,404*
28. Emily Bronte	1	Wuthering Heights, 1847	–

(*Continued*)

TABLE A1 | Continued

Author	Nbr. of texts	Example text in GLEC, year of publication	*GEPC text, length (nbr. of words)*
29. Ezra Pound	2	Certain Noble Plays of Japan, 1916	*Hugh Selwyn Mauberley, 1,181*
30. George A Henty	89	Under Drake's Flag, 1883	–
31. George Bernard Shaw	42	Pygmalion, 1912	–
32. George Eliot	13	Middlemarch, 1871	*How Lisa Loved the King, 2,702*
33. George Washington	1	State of the Union Addresses of George Washington, 1790	–
34. GK Chesterton	39	The Wisdom of Father Brown, 1914	*Complete Poems, 29,867*
35. Hamlin Garland	22	Money Magic, 1907	–
36. Harold Bindloss	43	Delilah of the Snows, 1907	–
37. Harriet EB Stowe	12	Uncle Tom's Cabin, 1852	–
38. Hector Hugh Munro	7	The Toys of Peace, 1919	–
39. Henry David Thoreau	9	Walden and on the Duty of Civil Disobedience, 1854	–
40. Henry James	72	The Golden Bowl, 1904	–
41. Henry Rider Haggard	52	Love Eternal, 1918	–
42. Herbert George Wells	51	The War of the Worlds, 1897	–
43. Herbert Spencer	4	The Philosophy of Style, 1880	–
44. Herman Melville	16	Moby Dick, 1851	*Poems, 19,088*
45. Howard Pyle	11	The Merry Adventures of Robin Hood, 1883	–
46. Isaac Asimov	1	Youth, 1952	–
47. Jack London	48	The Sea-Wolf, 1904	–
48. Jacob Abbott	47	William the Conqueror, 1849	–
49. James Bowker	1	Goblin Tales of Lancashire, 1878	–
50. James F Cooper	36	The Last of the Mohicans, 1826	–
51. James Joyce	4	Ulysses, 1922	*Chamber Music, 1,221*
52. James Matthew Barrie	23	Peter Pan, 1911	–
53. James Otis (Kaler)	27	Dick in the Desert, 1893	–
54. James Russell Lowell	11	Abraham Lincoln, 1890	*The Complete Poetical Works, 45,204*
55. Jane Austen	8	Emma, 1815	–
56. Jerome K Jerome	30	Three men in a Boat, 1898	–
57. John Bunyan	9	The Holy War, 1682	–
58. John Dryden	13	All for Love, 1678	*The Poetical Works, 80,667*
59. John Galsworthy	40	The Forsyte Saga, 1906–1921	–
60. John Keats	6	Endymion, 1818	*Poems, 36,408*
61. John Locke	3	An Essay Concerning Humane Understanding, 1689	–
62. John Maynard Keynes	1	The Economic Consequences of the Peace, 1919	–

(*Continued*)

TABLE A1 | Continued

Author	Nbr. of texts	Example text in GLEC, year of publication	*GEPC text, length (nbr. of words)*
63. John Morley	28	On Compromise, 1874	–
64. John Ruskin	42	A Joy For Ever, 1885	–
65. John Stuart Mill	11	Utilitarianism, 1861	–
66. Jonathan Swift	15	Gulliver's Travels, 1726	*The poems, 85,834*
67. Joseph Conrad	34	Lord Jim, 1899	–
68. Leigh Hunt	3	Stories From the Italian Poets/ With Lives of the Writers, 1835	*Captain Sword and Captain Pen, 2,260*
69. Lewis Carroll	14	Symbolic Logic, 1896	*Poems, 15,505*
70. Lord Byron	12	Fugitive Pieces, 1806	*Poetical Works, 207,977*
71. Lord Tennyson	10	Lady Clara Vere de Vere, 1842	*The Poems, 105,650*
72. Louisa May Alcott	34	Little Women, 1869	*Three Unpublished Poems, 386*
73. Lucy M Montgomery	17	Anne of Green Gables, 1908	–
74. Lyman Frank Baum	42	The Wonderful Wizard of Oz, 1900	–
75. Mark Twain	46	The Adventures of Tom Sawyer, 1876	–
76. Mary Shelley	5	Frankenstein, 1818	–
77. Michael Faraday	2	Experimental Researches in Electricity, 1839	–
78. Mary Stewart Daggett	2	Mariposilla, 1895	–
79. Nathaniel Hawthorne	88	The Scarlet Letter, 1850	–
80. O Henry	14	The Gift of the Magi, 1905	–
81. Oscar Wilde	25	The Picture of Dorian Gray, 1890	*Poems, 22089*
82. PB Shelley	7	Adonais, 1821	*The Complete Poetical Works, 165,242*
83. PG Wodehouse	35	A Damsel in Distress, 1919	–
84. Percival Lowell	2	The Soul of the Far East, 1896	–
85. Philip Kindred Dick	11	Mr. Spaceship, 1953	–
86. R M Ballantyne	88	The Red Eric, 1863	–
87. Rafael Sabatini	17	Scaramouche, 1921	–
88. Ralph Waldo Emerson	7	Nature, 1836	*Poems, 29,446*
89. Richard B Sheridan	5	Scarborough and the Critic, 1751	–
90. Robert Browning	7	Men and Women, 1855	*Poems, 35,732*
91. Robert Frost		A Boy's will, 1913	*Poems, 15,518*
92. Robert Hooke	1	Micrographia, 1665	–
93. Robert L Stevenson	79	A Childs Garden of Verses, 1885	*Poems, 33,755*
94. Robert Southey	3	The Life of Horatio Lord Nelson, 1798	*Poems, 23,857*
95. Rudyard Kipling	42	The Jungle Book, 1894	*Poems, 64,137*

(*Continued*)

TABLE A1 | Continued

Author	Nbr. of texts	Example text in GLEC, year of publication	*GEPC text, length (nbr. of words)*
96. Samuel T Coleridge	13	The Rime of the Ancient Mariner, 1798	*The Complete Poetical Works, 51,983*
97. Sinclair Lewis	7	Babbitt, 1922	–
98. Sir Arthur Conan Doyle	57	The Adventures of Sherlock Holmes, 1892	*Poems, 14,386*
99. Sir Francis Galton	3	Inquiries Into Human Faculty and its Development, 1883	–
100. Sir Humphry Davy	1	Consolations in Travel, 1830	–
101. Sir Isaac Newton	3	Opticks, 1704	–
102. Sir Joseph Dalton Hooker	1	Himalayan Journals, 1854	–
103. Sir Richard Francis Burton	11	The Land of Midian, 1877	–
104. Sir Walter Scott	35	Ivanhoe, 1820	*Poems, 46,846*
105. Sir Winston Churchill	4	The River War, 1899	–
106. Sir William Schwenck Gilbert	5	Songs of a Savoyard, 1890	*Poems, 31,138*
107. Stephen Leacock	15	Frenzied Fiction, 1917	–
108. TS Eliot	4	The Waste Land, 1922	*Poems, 4,661*
109. Thomas Carlyle	32	History of Friedrich II of Prussia, 1895	–
110. Thomas Crofton Croker	1	A Walk From London to Fulham, 1813	–
111. Thomas Hardy	26	Tess of the d'Urbervilles, 1891	*Poems, 62,756*
112. Thomas Henry Huxley	44	Darwinian Essays, 1893	–
113. Thomas Robert Malthus	4	An Essay on the Principle of Population, 1798	–
114. Thornton Waldo Burgess	31	Mrs. Peter Rabbit, 1902	–
115. Ulysses Grant	3	State of the Union Addresses, 1875	–
116. Virginia Woolf	4	Night and Day, 1919	–
117. Walt Whitman	5	Leaves of Grass, 1855	*Poems, 24,787*
118. Walter de la Mare	10	The Return, 1910	*Collected Poems, 15,765*
119. Washington Irving	17	The Legend of Sleepy Hollow, 1820	–
120. Wilkie Collins	32	Hide and Seek, 1854	–
121. William Blake	3	Songs of Innocence, 1789	*Poems, 4,439*
122. William Butler Yeats	24	In the Seven Woods, 1903	*Poems, 23,325*
123. William Dean Howells	84	Annie Kilburn, 1888	*Poems, 13,554*
124. William Ewart Gladstone	1	On Books and the Housing of Them, 1890	–
125. William Henry Hudson	13	The Purple Land, 1885	–
126. William J Long	8	Ways of Wood Folk, 1899	–
127. William M Thackeray	30	Barry Lyndon, 1844	*Ballads, 20,521*
128. William Penn	2	A Brief Account of the Rise and Progress of the People Called Quakers, 1698	–
129. William Shakespeare	38	Macbeth, 1623	*Sonnets, 8,721*
130. William Somerset Maugham	13	Of Human Bondage, 1915	–
131. William Wordsworth	7	I Wandered Lonely as a Cloud, 1807	*The Poetical Works, 116,683*
132. Winston Churchill (novelist)	13	The Inside of the Cup, 1913	–

word was assigned a value according to the number of graphemes belonging to the 10 rank sets. To control for word length, the sum of the values was divided by the number of graphemes per word. Thus, MEMORY would get a value of 9×2 [e o] + 2×5 [m] + 1×7 [r] + 1×8 [y = /i/] = 44/6 = 7.33, whereas SKUNK would get a value of 18/5 = 3.6. The final global sonority score of a poem is simply the mean of all word values in the poem. Of course, this simple additive model is only a first approximation, given the lack of any empirical data that would justify more complex models. Moreover, the fact that identical graphemes can have multiple context-dependent pronunciations in English like the/a/in "hAndbAll in the pArk" (Ziegler et al., 1997) is neglected in this first approximation which considers written, not spoken verbal materials.

B. Computing Word Similarity, Valence, and Arousal

Following upon an early unsupervised learning approach proposed by Turney and Littman (2003) and own previous theory-guided research (Westbury et al., 2014), I computed the lexical features *valence* and *arousal* on the basis of (taxonomy-based) semantic associations of a target word with a set of *labels*, i.e., keywords assumed to be prototypical for a certain affect, e.g., positive valence. The procedure for computing valence and arousal—implemented as a python script—was as follows. The script compared every target word with every word in the NLTK wordnet/WN database and computed the pairwise similarities (WNsim in Eq. A1 below, based on WN's path-similarity metric), summed and averaged them for each target word and then computed the difference between the mean for the positive and negative lists (for valence, not for arousal where the values were summed and averaged only):

$$\begin{aligned} &\mathrm{mean}[\mathrm{WNsim}(\mathit{word},\ \mathrm{label\_1pos}) \\ &\quad + \ldots + \mathrm{WNsim}(\mathit{word},\ \mathrm{label\_Npos})] \\ &\quad - \ \mathrm{mean}[\mathrm{WNsim}\ (\mathit{word},\ \mathrm{label\_1neg}) \\ &\quad + \ldots + \mathrm{WNsim}(\mathit{word},\ \mathrm{label\_Nneg})] \end{aligned} \tag{A1}$$

where label_1pos/1neg and label_Npos/Nneg are the first and last terms, respectively, in the valence lists given below.

The hit rates (i.e., overlap between words in the WN database and the present target words) were 80% for the Joyce poem and 77% for Eliot's, which can be considered as reliable (Jacobs and Kinder, 2017).

Label words for the computation of positive and negative valences, as well as arousal (for details, see Westbury et al., 2014, **Table 2**, row 2).

pos = ["contentment," "happiness," "pleasure," "pride," "relief," "satisfaction," "surprise"]
neg = ["disgust," "embarrassment," "fear," "sadness," "shame"]
aro = ["amusement," "anger," "contempt," "contentment," "disgust," "embarrassment," "excitement," "fear," "happiness," "interest," "pleasure," "relief," "sadness," "satisfaction"]