Linguistic **I**ssues in **L**anguage **T**echnology – **LiLT**
Submitted, October 2015

Distinguishing Voices in *The Waste Land* using Computational Stylistics

Julian Brooke
Adam Hammond
Graeme Hirst

Published by CSLI Publications

Distinguishing Voices in *The Waste Land* using Computational Stylistics

JULIAN BROOKE, *Department of Computing and Information Systems, University of Melbourne, jabrooke@unimelb.edu.au*
ADAM HAMMOND, *Department of English and Comparative Literature, San Diego State University, ahammond@mail.sdsu.edu*
GRAEME HIRST, *Department of Computer Science, University of Toronto, gh@cs.toronto.edu*

Abstract

T. S. Eliot's poem *The Waste Land* is a notoriously challenging example of modernist poetry, mixing the independent viewpoints of over ten distinct characters without any clear demarcation of which voice is speaking when. In this work, we apply unsupervised techniques in computational stylistics to distinguish the particular styles of these voices, offering a computer's perspective on longstanding debates in literary analysis. Our work includes a model for stylistic segmentation that looks for points of maximum stylistic variation, a *k*-means clustering model for detecting non-contiguous speech from the same voice, and a stylistic profiling approach which makes use of lexical resources built from a much larger collection of literary texts. Evaluating using an expert interpretation, we show clear progress in distinguishing the voices of *The Waste Land* as compared to appropriate baselines, and we also offer quantitative evidence both for and against that particular interpretation.

This work was carried out while the first author was at the University of Toronto.

LiLT Volume 12, Issue 2, October 2015.
Distinguishing Voices in ***The Waste Land*** *using Computational Stylistics.*

1 Introduction

Most work in automated stylistic analysis operates at the level of a text, assuming that a text is stylistically homogeneous. However, there are a number of instances where that assumption is unwarranted. One example is documents collaboratively created by multiple authors, in which contributors may, either inadvertently or deliberately (e.g., Wikipedia vandalism), create text which fails to form a stylistically coherent whole. Similarly, stylistic inconsistency might also arise when one of the 'contributors' is actually not one of the purported authors of the work at all – that is, in cases of plagiarism. More deliberate forms of stylistic dissonance include satire, which may first follow and then flout the stylistic norms of a genre, and much narrative literature, in which the author may give the speech or thought patterns of a particular character their own style distinct from that of the narrator. In this paper, we address this last source of heterogeneity in the context of the well-known poem *The Waste Land* by T. S. Eliot, which is often analyzed in terms of the distinct voices that appear throughout the text. The goal of the present work is to investigate whether computational stylistic analysis can distinguish these voices in ways that correspond to human interpretations, and also to explore whether our analysis can inform human interpretation, i.e., contribute to literary analysis.

First, we consider the problem of dividing the text into stylistically distinct segments. Our approach is inspired by research in topic segmentation (Hearst, 1994) and intrinsic plagiarism detection (Stamatatos, 2009), which is based on deriving a curve representing stylistic change, where the local maxima represent likely transition points. Notably, our curve represents an amalgamation of different stylistic metrics, including those that incorporate external (extrinsic) knowledge, e.g., vector representations based on larger corpus co-occurrence, which we show to be extremely useful. For development and initial testing we follow other work on stylistic inconsistency by using (artificially) mixed poems, but our main evaluation is on *The Waste Land* itself.

Next, we assume an initial segmentation and then try to create clusters corresponding to segments of *The Waste Land* which are spoken by the same voice, using the same features as the segmentation task. Of particular interest is the influence of the initial segmentation on the success of this downstream task.

Finally, we put aside the task-based approach of the first two sections and use an automatically generated but human-interpretable lexical stylistic resource to analyze the stylistic consistency of the major characters of the poem, as interpreted by our expert. Here, we consider the possibility that our gold standard could be wrong, and to settle the case we look closer at specific passages of the poem where discrepancies occur, and address them.

2 Related work

Poetry has been the subject of extensive computational analysis since the early days of literary and linguistic computing; see, e.g., (Beatie, 1967). Most of the research concerned either authorship attribution or analysis of metre, rhyme, and phonetic properties of the texts, but some work has studied the style, structure and content of poems with the aim of better understanding their qualities as literary texts. Among research that, like the present paper, looks at variation within a single text, Simonton (1990) found quantitative changes in lexical diversity and semantic classes of imagery across the components of Shakespeare's sonnets, and demonstrated correlations between some of these measures and judgments of the "aesthetic success" of individual sonnets. Duggan (1973) developed statistical measures of formulaic style to determine whether the eleventh-century epic poem *Chanson de Roland* manifests primarily an oral or a written style. Also related to our work, although it concerned a novel rather than a poem, is the paper by McKenna and Antonia (2001), who used principal component analysis of lexical frequency to discriminate different voices (dialogue, interior monologue, and narrative) and different narrative styles in sections of *Ulysses* by James Joyce.

One key property of the present work is the use of extrinsic lexical resources to analyze the style of literature. Kao and Jurafsky (2012), for instance, quantified various lexical aspects of poetry, including stylistic aspects such as abstractness, to distinguish professional and amateur writers of contemporary poetry. Voigt and Jurafsky (2013) investigated the usage of literary Chinese words in modern poetry. There has also been work in corpus linguistics that uses semantic tags to analyze characterization in literature, looking at *Romeo and Juliet* (Culpeper, 2009) and Virginia Woolf's *The Waves* (Balossi, 2014). More relevant to our stylistic interests is work by DeForest and Johnson (2000) using an automatically-built lexicon of Latinate words to distinguish pretentious characters in Jane Austen, and our own recent work with our six-style lexical model (see Section 6) that investigates how Virginia Woolf uses lexical choice in *To The Lighthouse* to differentiate the viewpoint narration of her characters with respect to their social background (Brooke et al., to appear).

More general work on identifying stylistic inconsistency includes that of Graham et al. (2005), who built artificial examples of style shift by concatenating different authors' Usenet postings. Feature sets for their neural network classifiers included standard textual features, frequencies of function words, punctuation and parts of speech, lexical entropy, and vocabulary richness. Guthrie (2008) presented some general methods of identifying stylistically anomalous segments using feature vector distance, and tested the effectiveness of his unsupervised method with a number of possible stylistic variations. He used features such as simple textual metrics (e.g., word and sentence length), readability measures, obscure vocabulary features, frequency rankings of function words (which were not found to be useful), and context analysis features from the General Inquirer dictionary. The most effective method ranked each segment according to the city-block distance of its feature vector to the feature vector of the textual complement (the union of all other segments in the text). Koppel et al. (2011) used a semi-supervised method to identify segments from two different books of the Bible artificially mixed into a single text. They first demonstrated that, in this context, preferred synonym use is a key stylistic feature that can serve as high-precision bootstrap for building a supervised SVM classifier on more general features (common words); they then used this classifier to provide an initial prediction for each verse and smooth the results over adjacent segments. The method crucially relied on properties of the King James Version translation of the text in order to identify synonym preferences. Finally, there is also stylometric work on distinguishing different authors within the same literary text using "rolling window" approaches somewhat similar to ours here (Eder, 2015), in particular those based on Burrows's (1987) classic approach, which also focuses on distributions of common words.

The identification of stylistic inconsistency or heterogeneity has received particular attention as a component of intrinsic plagiarism detection – the task of "identify[ing] potential plagiarism by analyzing a document with respect to undeclared changes in writing style" (Stein et al., 2011). A typical approach is to move a sliding window over the text looking for areas which are outliers with respect to the style of the rest of the text, or which differ markedly from other regions in word or character-trigram frequencies (Oberreuter et al., 2011, Kestemont et al., 2011). In particular, Stamatatos (2009) used a window to create a character trigram feature vector for each step through the text and then compared the vectors using a special distance metric to create a style change function whose maxima indicate points of interest (potential plagiarism).

With regards to the segmentation aspect of this work, topic segmentation is a similar problem that has been quite well explored. A common thread in this work is the importance of lexical cohesion, though a large number of competing models based on this concept have been proposed. One popular unsupervised approach is to identify the points in the text where a metric of lexical coherence is at a local minimum (Hearst, 1994, Galley et al., 2003). Malioutov and Barzilay (2006) also used a lexical coherence metric, but applied a graphical model where segmentations are graph cuts chosen to maximize coherence of sentences within a segment, and minimize coherence among sentences in different segments. Another class of approaches is based on a generative model of text, for instance HMMs (Blei and Moreno, 2001) and Bayesian topic modeling (Utiyama and Isahara, 2001, Eisenstein and Barzilay, 2008); in such approaches, the goal is to choose segment breaks that maximize the probability of generating the text, under the assumption that each segment has a different language model. Very recent relevant work applied a clustering approach, Affinity Propagation, to hierarchical topic segmentation within a novel (Kazantseva and Szpakowicz, 2014).

Though in practice making sense of viewpoint and narrative in a poem such as *The Waste Land* is a very different task than it would be in the context of prose fiction – and this is particularly true in light of the stylistic focus of this work – applications of computational techniques in identifying the narrative structure of novels is also relevant, albeit tangentially. Consider, for instance, Wiebe's (1994) rule-based system for identifying change in character viewpoint, and Wallace's (2012) Bayesian approach to clustering narrative threads.

3 *The Waste Land*

3.1 Background

T. S. Eliot (1888-1965), recipient of the 1948 Nobel Prize for Literature, is among the most important twentieth-century writers in the English language. Born in St. Louis, he studied literature and philosophy at Harvard and Oxford before settling in London. He published his first poem, "The Love Song of J. Alfred Prufrock", in 1915. In the years that followed, he published several volumes of poetry, worked as a literary critic as well as a banker, founded the influential literary periodical *The Criterion*, and turned increasingly toward playwriting, winning a Tony Award in 1950. Though he worked in many forms, he is best remembered today for his early poetry, of which *The Waste Land* (1922) is his most important single work.

The poem was composed during a period of personal distress in Eliot's life: on the verge of a nervous breakdown brought about by marital difficulties, he headed to a Swiss sanatorium in November 1921, returning two months later with a manuscript of the poem. *The Waste Land* deals with themes of spiritual and cultural death, offering little hope of rebirth or salvation. It is notable for its formal experimentations: it features a disjunctive structure with rapid and unmarked transitions between scenes and character voices; it makes numerous uncited references to a wide array of cultural materials, from high culture to low culture, often presenting them in their original language; and it features a syncopated rhythm borrowed from jazz. The poem is divided into five parts; in total it is 433 lines long, and contains 3533 tokens, not including the headings.

A prominent debate among scholars of *The Waste Land* concerns whether a single speaker's voice predominates in the poem (Bedient, 1986), or whether the poem should be regarded instead as dramatic or operatic in structure, composed of about twelve different voices independent of a single speaker (Cooper, 1987). Several facts support the latter argument. At the time *The Waste Land* was published, Eliot was beginning to move away from lyric poetry (compositions with a single speaking voice) toward drama (compositions consisting of numerous differentiated voices). Indeed, in his notes to *The Waste Land*, Eliot supports the view of the poem as a dramatic composition by referring to its "characters" and "personage[s]". Eliot's working title, *He Do the Police in Different Voices*, further supports this argument. Another clue is provided in the poem's reference in lines 127-130 to the 1912 popular song "The Shakespearean Rag" – for *The Waste Land* itself functions much like a "ragtime song" (or "rag"), a genre defined by its combination of many "scraps" of culture and many "tissues" of fragmentary voices into a single composition (Sigg, 1994).

3.2 The Voices

Among the most distinctive voices in *The Waste Land* is the woman who speaks at the end of its second section:

> I can't help it, she said, pulling a long face,
> It's them pills I took, to bring it off, she said
> [158-159]

"Woman in Bar", as we will refer to her, has a chatty tone and lower-class speech patterns that distinguish her voice from many others in the poem, such as the traditionally poetic voice of a narrator ("Tiresias") that recurs many times in the poem:

> Above the antique mantel was displayed
> As though a window gave upon the sylvan scene
> The change of Philomel
> [97-99]

There are nonetheless other educated, poetic voices in *The Waste Land*. Near the beginning of the poem, for example, there is a rhythmic voice (the "Hellfire Preacher") that quotes the Old Testament, often with anger or vehemence:

What are the roots that clutch, what branches grow
Out of this stony rubbish? Son of man,
You cannot say, or guess, for you know only
A heap of broken images
[19-22]

Likewise, toward the end of the poem, we hear an educated, polyglot voice ("Crazy Prufrock", so named because he bears similarities to the protagonist of Eliot's earlier poem "The Love Song of J. Alfred Prufrock") that quotes a wide variety of literary sources, yet also seems to be suffering from mental illness:

I sat upon the shore
Fishing, with the arid plain behind me
Shall I at least set my lands in order?
London Bridge is falling down falling down falling down
[424427]

Another voice,"Nervous One", is similarly deranged, intense, and intellectual, but is clearly identified as a female:

"What shall I do now? What shall I do?"
"I shall rush out as I am, and walk the street"
"With my hair down, so. What shall we do to-morrow?"
"What shall we ever do?"
[131-134]

Certain voices evoke the media environment of the early twentieth century, such as "Intrepid Reporter", who speaks in the clipped, strident tone of a newsreel announcer:

Madame Sosostris, famous clairvoyante,
Had a bad cold, nevertheless
Is known to be the wisest woman in Europe
[43-45]

The Tarot-reading Madame Sosostris speaks in the slightly ungrammatical voice of a non-native speaker:

If you see dear Mrs. Equitone,
Tell her I bring the horoscope myself
[57-59]

More subtly marked voices included that of Marie, an older aristocratic woman who speaks German and is given to nostalgic remembrances of her past:

Summer surprised us, coming over the Starnbergersee
With a shower of rain; we stopped in the colonnade,
And went on in sunlight, into the Hofgarten,
And drank coffee, and talked for an hour.

> [8-11]

Another female voice, “The Typist”, is distraught and unhappy, yet resigned and sober in her tone:

> “Well now that’s done: and I’m glad it’s over.”
> [252]

For passages that seem to come not from a single speaker, but from the voice of a larger group or the community itself, we have adapted the convention of Greek tragedy and labeled this voice “The Chorus”:

> Elizabeth and Leicester
> Beating oars
> The stern was formed
> A gilded shell
> Red and gold
> The brisk swell
> Rippled both shores
> Southwest wind
> Carried down stream
> The peal of bells
> [279-288]

While the stylistic contrasts between these and other, briefly appearing voices are apparent to many readers, Eliot does not explicitly mark the transitions between them.

4 Segmenting Voices

4.1 Stylistic change curves

Many popular text segmentation methods depend crucially on a base textual unit (often a sentence) which can be reliably classified or compared to others.[1] In the context of stylistic analysis of poetry such as the *The Waste Land*, however, we do not want to commit to a particular unit beyond the word token, since any unit large enough to provide a reliable stylistic footprint (e.g., a stanza) will also render many of the most interesting voice switches outside the scope of our model (see examples in Section 4.4). Generative models, which use a bag-of-words assumption, have a very different problem: in their standard form, they can capture *only* lexical cohesion, which is not the (primary) focus of stylistic analysis. In particular, we wish to segment using information that goes beyond the distribution of words in the text being segmented. The model for stylistic segmentation we propose here is related to Hearst’s (1994) TextTiling technique and to Stamatatos’s (2009) style change function, but our model is generalized so that it applies to any numeric metric (feature)

[1] Section 4 is adapted from our earlier published work (Brooke et al., 2012).

that is defined over a span; importantly, style change curves represent the change of a set of very diverse features.

Our goal is to find the precise points in the text where a stylistic change (a voice switch) occurs. To do this, we calculate, for each token in the text, a measure of stylistic change which corresponds to the distance of feature vectors derived from a fixed-length span on either side of that point. That is, if $\mathbf{v}_{ij}$ represents a feature vector derived from the tokens between (inclusive) indices i and j, then the stylistic change c_i at point i for a span (window) of size w is:

$$c_i = Dist(\mathbf{v}_{(i-w)(i-1)}, \mathbf{v}_{i(i+w-1)})$$

This function is not defined within w tokens of the edge of the text, and we generally ignore the possibility of breaks within these (unreliable) spans. Possible distance metrics include cosine distance, Euclidean (L_2) distance, and city-block (L_1) distance. In his study, Guthrie (2008) found best results with city-block distance, and that is what we will primarily use here. For some pair of feature vectors **f** and **g**, both of length l, city block distance is defined as:

$$Dist(\mathbf{f},\mathbf{g}) = \sum_{i=0}^{l} |\mathbf{f}_i - \mathbf{g}_i|$$

The feature vector can consist of any features that are defined over a span; one important step, however, is to normalize each feature (here, to a mean of 0 and a standard deviation of 1), so that different scaling of features does not result in particular features having an undue influence on the stylistic change metric. That is, if some feature is originally measured to be f_i in the span i to $i + w - 1$, then its normalized version f_i' (included in $\mathbf{v}_{i(i+w-1)}$) is:

$$f_i' = \frac{f_i - \overline{f}}{\sigma_f}$$

The local maxima of c represent our best predictions for the stylistic breaks within a text. Stylistic change curves, however, are not well behaved; they may contain numerous spurious local maxima if a local maximum is defined simply as a higher value between two lower ones. We can narrow our definition by requiring that the local maximum be maximal within some window w'. That is, our breakpoints are those points i where, for all points j in the span $x - w', x + w'$, it is the case that $c_i > c_j$. As it happens, $w' = w / 2$ is a fairly good choice for our purposes, creating spans no smaller than the smoothed window, though w' can be lowered to increase breaks, or increased to limit them. The absolute height of the curve at each local minimum offers a secondary way of ranking (and eliminating) potential breakpoints, if more precision is required; however, in our task here the breaks are fairly regular but often subtle, so focusing only on the largest stylistic shifts is not necessarily desirable.

4.2 Features

The features we explore for this task fall roughly into two categories: surface and extrinsic. The distinction is not entirely clear-cut, but we wish to distinguish features that use the basic properties of the words or their part of speech (PoS), which have traditionally been the focus of automated stylistic analysis, from features which rely heavily on external lexical information, for instance word sentiment and, in particular, vector space representations, which are more novel for this task. Our features are listed below.

Word length A common textual statistic in register and readability studies. Readability, in turn, has been used for plagiarism detection (Stein et al., 2011), and related metrics were consistently among the best in (Guthrie, 2008).

Syllable count Syllable count is a reasonably good predictor of the difficulty of a vocabulary, and is used in some readability metrics.

Punctuation frequency The presence or absence of punctuation such as commas, colons, and semicolons can be a very good indicator of style. We also include periods, which offer a measure of sentence length.

Line breaks Our only poetry-specific feature; we count the number of times the end of a line appears in the span. More or fewer line breaks (that is, longer or shorter lines) can vary the rhythm of the text, and thus its overall feel.

Parts of speech Lexical categories can indicate, for instance, the degree of nominalization, which is a key stylistic variable (Biber, 1988). We collect statistics for the four main lexical categories (noun, verb, adjective, adverb) as well as prepositions, determiners and proper nouns.

Pronouns We count the frequency of first-, second- and third-person pronouns, which can indicate the interactiveness and narrative character of a text (Biber, 1988).

Verb tense Past tense is often preferred in narratives, whereas present tense can give a sense of immediacy.

Type-token ratio A standard measure of lexical diversity.

Lexical density Lexical density is the ratio of the count of tokens of the four substantive parts of speech to the count of all tokens.

Contextuality measure The contextuality measure of Heylighen and Dewaele (2002) is based on PoS tags (e.g., nouns decrease contextuality, while verbs increase it), and has been used to distinguish formality in collaboratively built encyclopedias (Emigh and Herring, 2005).

Dynamic In addition to the hand-picked features above, we test dynamic inclusion of words and character trigrams that are common in the text being analyzed, particularly those not evenly distributed throughout the text (we exclude punctuation). To measure the latter, we define *clumpiness* as the square root of the index of dispersion or variance-to-mean ratio (Cox and Lewis, 1966) of the (text-length) normalized differences between successive occurrences of a feature, including (importantly) the difference between the first index of the text and the first occurrence of the feature as well as the last occurrence and the last index; the measure varies between 0 and 1, with 0 indicating a perfectly even distribution. We test with the top *n* features based on the ranking of the product of the feature's frequency in the text (*tf*) or product of the frequency and its clumpiness (*tf-cl*); this is similar to a *tf-idf* weight.

Next, we turn to the extrinsic features. For those lexicons which include only lemmatized forms, the words are lemmatized before their values are retrieved.

Percent of words in Dale-Chall Word List A list of 3000 basic words that is used in the Dale-Chall Readability metric (Dale and Chall, 1995).

Average unigram count in 1T Corpus Another metric of whether a word is commonly used. We use the unigram counts in the 1T 5-gram Corpus (Brants and Franz, 2006); each token in the span is assigned a (type) count, and the average over the span is taken.[2] Here and below, if a word is not included, it is given a zero.

Sentiment polarity The positive or negative stance of a span could be viewed as a stylistic variable. We test two lexicons, a hand-built lexicon for the SO-CAL sentiment analysis system which has shown superior performance in lexicon-based sentiment analysis (Taboada et al., 2011), and SentiWordNet (SWN), a high-coverage automatic lexicon built from WordNet (Baccianella et al., 2010). The polarity of each word over the span is averaged.

Sentiment extremity Both sentiment lexicons provide a measure of the degree to which a word is positive or negative. Instead of averaging the sentiment scores, we average the absolute values of those scores. High values reflect strong sentiment in the text but do not indicate whether that sentiment is positive or negative.

[2] The 1T Corpus reflects the English of the World Wide Web in 2005–06, so its unigram frequencies for the vocabulary of *The Waste Land* will be only an approximation, albeit a close one in most cases, to those of the period, 1921–22, in which the poem was written.

Formality Average formality score, using a lexicon of formality (Brooke et al., 2010) built using latent semantic analysis (LSA) (Landauer and Dumais, 1997).

Dynamic General Inquirer The General Inquirer dictionary (Stone et al., 1966), which was used for stylistic inconsistency detection by Guthrie (2008), includes 182 content analysis tags, many of which are relevant to style; we remove the two polarity tags already part of the SO-CAL dictionary, and select others dynamically using our *tf-cl* metric.

LSA vector features Brooke et al. (2010) have posited that, in corpora that are highly diverse in register or genre, the lowest dimensions of word vectors derived using LSA (or other dimensionality reduction techniques) often reflect stylistic concerns; they found that using the first 20 dimensions to build their formality lexicon gave the best results in a near-synonym evaluation. Early work by Biber (1988) on the Brown Corpus using a related technique (factor analysis) resulted in discovery of several identifiable dimensions of register. Here, we investigate direct use of these LSA-derived vectors, with each of the first 20 dimensions corresponding to a separate feature. We test with vectors derived from the word-document matrix of the ICWSM 2009 blog dataset (Burton et al., 2009) which includes 1.3 billion tokens, and also from the BNC (Burnard, 2000), which has 100 million tokens.[3] The length of the vector depends greatly on the frequency of the word; since this is being accounted for elsewhere, we normalize each vector to the unit circle.

4.3 Evaluation method

Metrics

To evaluate our method, we apply standard topic segmentation metrics, comparing the segmentation boundaries to a gold-standard reference. The measure P_k, proposed by Beeferman et al. (1999), uses a probe window of length k set to half the average length of a segment; the window slides over the text, and counts the number of instances where a unit (in our case, a token) at one edge of the window was predicted to be in the same segment as a unit at the other edge, but in fact is not; or was predicted not to be in the same segment, but in fact (according to the reference) is. This count is normalized by the total number of tests to get a score between 0 and 1, with 0 being a perfect score (the lower, the better). If N is the number of units in the text, $b(r_i, r_{i+k})$ is the number of boundaries between points i and $i+k$ in the reference segmentation r, and $b(h_i, h_{i+k})$ is the number of boundaries between points i and $i+k$ in

[3] As the 1T corpus, these are modern corpora, and therefore word usage will differ somewhat from usage in 1922.

the hypothesized or predicted segmentation h, then

$$P_k(r,h) = \frac{1}{N-k}\sum_{i=1}^{N-k} \mathbb{1}((b(r_i, r_{i+k}) = 0 \text{ and } b(h_i, h_{i+k}) \neq 0) \text{ or } \\ (b(r_i, r_{i+k}) \neq 0 \text{ and } b(h_i, h_{i+k}) = 0)),$$

where $\mathbb{1}$ is the indicator function that returns 1 if its boolean-expression argument is true and 0 otherwise. Pevzner and Hearst (2002) criticize this metric because it penalizes false positives and false negatives differently and sometimes fails to penalize false positives altogether; their metric, *WindowDiff* (WD), solves these problems by counting an error whenever there is a difference between the number of segments in the prediction as compared to the reference. It is given by

$$WD(r,h) = \frac{1}{N-k}\sum_{i=1}^{N-k} \mathbb{1}(b(r_i, r_{i+k}) \neq b(h_i, h_{i+k})).$$

Relatively recent work in topic segmentation (Eisenstein and Barzilay, 2008) continues to use both metrics, so we also present both here.

During initial testing, we noted a shortcoming of both these metrics: all else being equal, they will usually prefer a system which predicts fewer breaks; in fact, a system that predicts no breaks at all can score less than 0.3 (a very competitive result both here and in topic segmentation) if the variation of the true segment size is reasonably high, which it is in *The Waste Land*. This is problematic because we do not want to be trivially ‘improving’ simply by moving towards a model that is too cautious to guess anything at all. We did not throw these metrics out, since they are otherwise very carefully designed to give an overall sense of segmentation quality, or create a new all-in-one metric without this problem, which we judged as being beyond the scope of this work. Instead, we created a rather crude recall-focused metric, which we call BD (break difference), which sums all the distances, calculated as fractions of the entire text, between each true break and the nearest predicted break. This metric is more seriously flawed than the other two, because it can be trivially made 0 (the best score) by guessing a break everywhere. However, when the two primary metrics improve (decrease) without a corresponding increase in BD, we can be fully confident that, with regards to approaching the gold standard, the new segmentation is actually better. Fortunately, this is true for most of the results presented here.

Baselines

We compare our method to the following baselines:

Random selection We randomly select boundaries, using the same number of boundaries in the reference. We average over 50 runs.

Evenly spaced We put boundaries at equally spaced points in the text, using the same number of boundaries as the reference.

Random feature We use our stylistic change curve method with a single feature which is created by assigning a uniform random value to each token and averaging across the span. Again, we average the score over 50 runs.

There are other possible baselines for *The Waste Land* that make use of text formatting (e.g., stanza boundaries), but for comparison in this context we only include baselines which can be derived from the same stream-of-tokens representation that our model uses, and which are also appropriate for our experiment involving artificially created poems.

4.4 Experiments

Artificial poems

Our main interest is *The Waste Land*. It is, however, prudent to develop our method, i.e., conduct an initial investigation of our method, including parameters and features, using a separate corpus. We do this by building artificial mixed-style poems by combining stylistically distinct poems by different authors, as others have done with prose.

Our set of twelve poems (or other verse) used for this evaluation was selected by one of the authors, an English literature expert, to reflect the stylistic range and influences of poetry at the beginning of the twentieth century, and on *The Waste Land* in particular.[4] The longest of these poems is 1291 tokens and the shortest is just 90 tokens (though 10 of the 12 have at least 300 tokens); the average length is 501 tokens. The titles were removed, and each poem was tagged by an automatic PoS tagger (Schmid, 1995). Koppel et al. (2011) built their composite version of two books of the Bible by choosing, at each step, a random span length (from a uniform distribution) to include from one of the two books being mixed, and then a span from the other, until all the text in both books had been included. Our method is similar, except that we first randomly select six poems to include in the particular mixed text, and at each step we randomly select one of poems, reselecting if the poem has been used up or the remaining length is below our lower bound. For our first experiment, we set a lower bound of 100 tokens and an upper bound of 200 tokens

[4] The specific poems used were: “September 1, 1939” by W. H. Auden, “Wagner” by Rupert Brooke, “The Love Song of J. Alfred Prufrock” by T. S. Eliot, “Ballad of Another Ophelia” by D. H. Lawrence, “Giovanni Franchi” by Mina Loy, “Strange Meeting” by Wilfred Owen, “How Should I Your True Love Know?” from *Hamlet* by William Shakespeare, “Not Waving But Drowning” by Stevie Smith, “Epithalamion” by Edmund Spenser, “Before the Beginning of Years” from *Atalanta in Calydon* by Algernon Charles Swinburne, “The Coming of Arthur” from *The Idylls of the King* by Alfred, Lord Tennyson, “A Saint About to Fall” by Dylan Thomas.

for each span; although this gives a higher average span length than that of *The Waste Land*, our first goal is to test whether our method works in the (ideal) condition where the feature vectors at the breakpoint generally represent spans which are purely one poem or another for a reasonably high *w* (100). We create 50 texts using this method. In addition to testing each individual feature, we test several combinations of features (all features, all surface features, all extrinsic features), and present the best results for greedy feature removal, starting with all features (excluding dynamic ones) and choosing features to remove in order to minimize the sum of the three metrics.

The Feature Sets section of Table 1 gives the individual feature results on segmenting the artificially combined poems. Using any of the features alone is better than our baselines, though some of the metrics (in particular type-token ratio) are only a slight improvement. Line breaks are obviously quite useful in the context of poetry (though the WD score is high, suggesting a precision/recall trade-off), but so are more typical stylistic features such as the distribution of basic lexical categories and punctuation. The unigram count and formality score are otherwise the best two individual features. The sentiment-based features performed more modestly, though the extremeness of polarity was useful when paired with the coverage of SentiWordNet. Among the larger feature sets, the General Inquirer was the least useful, though more effective than any of the individual features, while dynamic word and character trigrams did better, and the ICWSM LSA vectors better still; the difference in size between the ICWSM and BNC is obviously key to the performance difference here. In general using our *tf-cl* metric was better than *tf* alone.

When we combine the different feature types, we see that extrinsic features have a slight edge over the surface features, but the two do complement each other to some degree. Although the GI and dynamic feature sets do well individually, they do not combine well with other features in this unsupervised setting, and our best results do not include them. The greedy feature selector removed four LSA dimensions, type-token ratio, prepositions, second-person pronouns, adverbs and verbs to get our best result. Our choice of *w* to be the largest fully-reliable size (100) seems to be a good one, as is our use of city-block distance rather than the alternatives. Overall, the metrics we are using for evaluation suggest that we are roughly halfway to perfect segmentation.

Configuration	**Metrics**		
	WD	P_k	BD
Baselines			
Random breaks	0.532	0.465	0.465
Even spread	0.498	0.490	0.238
Random feature	0.507	0.494	0.212
Feature sets			
Word length	0.418	0.405	0.185
Syllable length	0.431	0.419	0.194
Punctuation	0.412	0.401	0.183
Line breaks	0.390	0.377	0.200
Lexical category	0.414	0.402	0.177
Pronouns	0.444	0.432	0.213
Verb tense	0.444	0.433	0.202
Lexical density	0.445	0.433	0.192
Contextuality	0.462	0.450	0.202
Type-Token ratio	0.494	0.481	0.204
Dynamic (*tf*, *n*=50)	0.399	0.386	0.161
Dynamic (*tf-cl*, 50)	0.385	0.373	0.168
Dynamic (*tf-cl*, 500)	0.337	0.323	0.165
Dale-Chall	0.483	0.471	0.202
Count in 1T	0.424	0.414	0.193
Polarity (SO-CAL)	0.466	0.487	0.209
Polarity (SWN)	0.490	0.478	0.221
Extremity (SO-CAL)	0.450	0.438	0.199
Extremity (SWN)	0.426	0.415	0.182
Formality	0.409	0.397	0.184
All LSA (ICWSM)	0.319	0.307	0.134
All LSA (BNC)	0.364	0.352	0.159
GI (*tf*, *n*=5)	0.486	0.472	0.201
GI (*tf-cl*, 5)	0.449	0.438	0.196
GI (*tf-cl*, 50)	0.384	0.373	0.164
Combinations			
Surface	0.316	0.304	0.150
Extrinsic	0.314	0.301	0.124
All	0.285	0.274	0.128
All w/o GI, dynamic	0.272	0.259	0.102
All greedy (Best)	**0.253**	**0.242**	**0.099**
Best, *w*=150	0.289	0.289	0.158
Best, *w*=50	0.338	0.321	0.109
Best, Diff=euclidean	0.258	0.247	0.102
Best, Diff=cosine	0.274	0.263	0.145

TABLE 1 Segmentation accuracy in artificial poems. Lower values are better.

The Waste Land

In order to evaluate our method on *The Waste Land*, we first created a gold-standard voice switch segmentation.[5] Our gold standard represents an amalgamation, by one of the authors, of several sources of information. First, we enlisted a class of 140 undergraduates in an English literature course to segment the poem into voices using their own intuitions. Second, our English literature expert listened to the six readings of the poem included on *The Waste Land* app (Touch Press LLP, 2011), including two readings by T. S. Eliot, and noted places where the reader's voice seemed to change. Next, versions of the poem were prepared which aggregated the voice-switches indicated by students and readers, on a token-by-token basis. Examples of this output are provided below, first for students and then for readers, where (VS: n) indicates *n* voice switches.

> *Students:*
> And when we were children, staying at the archduke's,(VS: 2)
> My cousin's,(VS: 1) he took me out on a sled,(VS: 1)
> And I was frightened.(VS: 11) He said,(VS: 27) Marie,
> Marie, hold on tight.(VS: 30) And down we went.(VS: 7)
> In the mountains,(VS: 1) there you feel free.(VS: 7)
> I read, much of the night, and go south in the winter.(VS: 68)
>
> What are the roots that clutch, what branches grow
> Out of this stony rubbish?(VS: 14) Son of man,(VS: 1)
> You cannot say, or guess, for you know only
> A heap of broken images(VS: 1)

> *Readers:*
> And when we were children, staying at the archduke's,
> My cousin's, he took me out on a sled,
> And I was frightened. He said,(VS: 3) Marie,
> Marie, hold on tight.(VS: 3) And down we went.(VS: 1)
> In the mountains, there you feel free.
> I read, much of the night, and go south in the winter.(VS: 5)
>
> What are the roots that clutch, what branches grow
> Out of this stony rubbish? Son of man,
> You cannot say, or guess, for you know only
> A heap of broken images

[5] The full gold standard can be browsed at http://hedothepolice.org/class/read.html. For a more machine-readable version, see http://www.cs.toronto.edu/∼jbrooke/wasteland_annotations.zip.

While no hard thresholds were established, the expert generally proceeded by investigating points where more than 20 students identified a voice switch, and compared the results with those of readers. Ignoring cases of reported speech (in this passage, “Marie, / Marie, hold on tight”, which is explicitly introduced as dialogue from within a single voice, not as a point of transition between voices), the expert marked a switch in the gold standard when the findings of the students agreed with those of the readers and his own reading of the poem. In the passage above, a single gold-standard switch was recorded: at the stanza break between “south in the winter” and “What are the roots”.

The second step in establishing the gold standard was to cluster the segmented poem into individual voices, an annotation that we make use of in Section 5. This was performed in an entirely qualitative manner, drawing on the expert’s knowledge of the poem as well as focusing on patterns of repetition and allusion among voices (for instance, if a distinctive phrase was repeated in two segments, or a poem was quoted in two segments, these were more likely to be clustered together).

We created two versions of the poem for evaluation. From both versions, we removed everything but the main body of the text (i.e., the prologue, dedication, title, and section titles), since these are not produced by voices in the poem. The *full* version contains all the other text (a total of 68 voice switches), but our *abridged* version involved removing all segments (and the corresponding voice switches, when appropriate) that have 20 or fewer tokens (including the 10 segments entirely in a language other than English); this reduces the number of voice switches to 28 (the token count is 3179). This version allows us to focus on the segmentation for which our method has a reasonable chance of succeeding and ignore the segmentation of non-English spans, which is relatively trivial but yet potentially confounding. We use $w = 50$ for the full version, since there are almost twice as many breaks as in the abridged version (and our artificially generated texts).

Our results for *The Waste Land* are presented in Table 2. Notably, in this evaluation, we do not investigate the usefulness of individual features or attempt to fully optimize our solution using this text. Our goal is to see if a general stylistic segmentation system, developed on artificial texts, can be applied successfully to the task of segmenting an actual stylistically diverse poem. The answer is yes. Although the task is clearly more difficult, the results for the system are well above the baseline, particularly for the abridged version. One thing to note is that using the features greedily selected for the artificial system (instead of just all features) appears to hinder, rather than help; this suggests a supervised approach might not be effective. The General Inquirer is too unreliable

Configuration	**Metrics**		
	WD	P_k	BD
Full text			
Baselines			
Random breaks	0.517	0.459	0.480
Even spread	0.559	0.498	0.245
Random feature	0.529	0.478	0.314
System (*w*=50)			
Table 1 Best	0.458	0.401	0.264
GI	0.508	0.462	0.339
Dynamic	0.467	0.397	0.257
LSA (ICWSM)	0.462	0.399	0.280
All w/o GI	**0.448**	0.395	0.305
All w/o dynamic, GI	0.456	**0.394**	**0.228**
Abridged text			
Baselines			
Random breaks	0.524	0.478	0.448
Even spread	0.573	0.549	0.266
Random feature	0.525	0.505	0.298
System (*w*=100)			
Table 1 Best	0.370	0.341	0.250
GI	0.510	0.492	0.353
Dynamic	0.415	0.393	0.274
LSA (ICWSM)	0.411	0.390	0.272
All w/o GI	0.379	0.354	0.241
All w/o dynamic, GI	**0.345**	**0.311**	**0.208**

TABLE 2 Segmentation accuracy in *The Waste Land*. Lower is better.

to be useful here, whereas the dynamic word and trigram features continue to do fairly well, but they do not improve the performance of the rest of the features combined. Once again the LSA features seem to play a central role in this success.

We manually compared predicted switches with real switches and found that there were several instances (corresponding to very clear voice switches in the text) which were nearly perfect. We can see this result very clearly in Figure 1, which shows that the local maximum of our change curve in the abridged version of the text often corresponds exactly to a real switch. Moreover, the model did tend to predict more switches in sections with numerous real switches, though these predictions were often fewer than the gold standard and out of sync (because the sampling

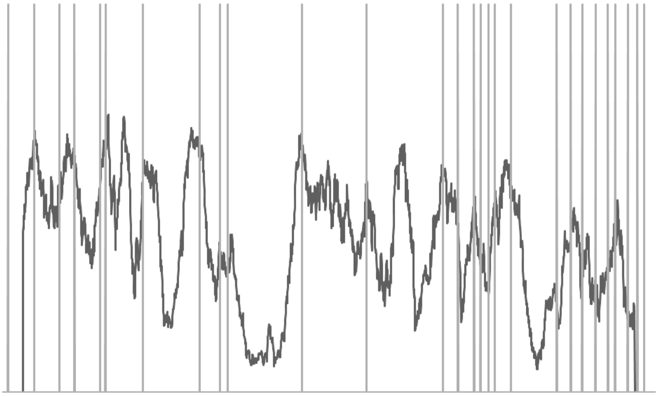

FIGURE 1 Stylistic change curve over the abridged version of *The Waste Land*. Vertical gray lines are gold-standard voice switches.

windows never consisted of a pure style). This is visible in Figure 1 in the last third of the poem, where the change curve is somewhat erratic in those areas with many breaks.

Though there were many instances where the model agreed with the expert annotation, and many others where the automatic stylistic segmentation failed to pick on obvious (to a human) discourse cues, there were a few instances where the automated output made a choice that was different but nonetheless interesting from the perspective of literary analysis. One example comes near the beginning of the poem, where there is a transition from the dark introductory musings of Tiresias (e.g., “April is the cruelest month”) to the viewpoint of Marie (e.g., “Summer surprised us, coming over the Starnbergersee”). Both segmentations had a break, but the automated stylistic annotation placed it several lines later than the gold standard – an interpretation we found was quite defensible and, in retrospect, perhaps even preferable; one of the expert readers on the Waste Land app made the same choice. In another case, our model inserted an extra break in a stanza presumed to belong entirely to a single character (Crazy Prufrock). This break allows for a more stylistically parsimonious interpretation of the stanza, which begins with a phrase that is a telltale indicator of Prufrock but, after the automatically determined break, does indeed not sound much like Prufrock. In light of these examples, we would argue that an automated stylistic segmentation model, rather than simply mimicking a particular human interpretation, can offer an intriguing alternative viewpoint on the voice switches in the poem. For interested readers, the human and computer voice switches for *The Waste Land* can be compared on the project website.[6]

5 Clustering Voices

5.1 Method

Our approach to voice identification in *The Waste Land* consists first of identifying the boundaries of voice spans, as outlined in section 4.[7] Then, given a segmentation of the text, we consider each span as a data point in a clustering problem. The elements of the vector correspond to the best feature set from the segmentation task, with the rationale that features which were useful for detecting changes in style should also be useful for identifying stylistic similarities.

For clustering, we use a slightly modified version of the popular *k*-means algorithm (MacQueen, 1967). Briefly, *k*-means assigns points to a cluster based on their proximity to the *k* cluster centroids, which are initialized to randomly chosen points from the data and then iteratively refined until convergence, which in our case was defined as a change of less than 0.0001 in the position of each centroid during one iteration.[8] Our version of *k*-means is distinct in two ways. First, it uses a weighted centroid where the influence of each point is based on the token length of the underlying span, i.e., short (unreliable) spans which fall into the range of some centroid will have less effect on the location of the centroid than larger spans. Second, we use a city-block (L_1) distance function rather than standard Euclidean (L_2) distance function; in the segmentation task, we found that city-block (L_1) distance was preferred, a result which is in line with other work in stylistic inconsistency detection (Guthrie, 2008). Though it would be interesting to see if a good *k* could be estimated independently, for our purposes here we set *k* to be the known number of speakers in our gold standard.

5.2 Evaluation

We evaluate our clusters by comparing them to a gold standard annotation. There are various metrics for extrinsic cluster evaluation; Amigó et al. (2009) review various options and select the BCubed precision and recall metrics (Bagga and Baldwin, 1998) as having all of a set of key desirable properties. BCubed precision is a calculation of the fraction of item pairs in the same cluster which are also in the same category,

[6] http://hedothepolice.org/computer/read.html

[7] Section 5 is adapted from our earlier published work (Brooke et al., 2013).

[8] Occasionally, there was no convergence, at which point we arbitrarily halted the process after 100 iterations.

whereas BCubed recall is the fraction of item pairs in the same category which are also in the same cluster. The harmonic mean of these two metrics is BCubed F-score. Typically, the “items” are exactly what has been clustered, but this is problematic in our case, because we wish to compare methods which have different segmentations and thus the vectors that are being clustered cannot be directly compared. Instead, we calculate the BCubed measures at the level of the token; that is, for the purposes of measuring performance we act as if we had clustered each token individually, instead of the spans of tokens actually used.

As in Section 4.4, our first evaluation is against a set of 20 artificially generated “poems”; our method for creating these poems is the same as in Section 4.4. Again, the idea is to allow us to evaluate our method in more ideal circumstances, i.e., when there are very distinct voices corresponding to different poets, and the voice spans tend to be fairly long. Our second evaluation is of *The Waste Land* itself, using the (unabridged) gold-standard annotation discussed earlier in Section 4.4.

We consider three segmentations: the segmentation of our gold standard (Gold), the segmentation predicted by our segmentation model (Automatic), and a segmentation which consists of equal-length spans (Even), with the same number of spans as in the gold standard. The Even segmentation should be viewed as the baseline for segmentation, and the Gold segmentation an “oracle” representing an upper bound on segmentation performance. For the automatic segmentation model, we use the best settings from Section 4.4. We also compare three possible clusterings for each segmentation: no clustering at all (Initial), that is, we assume that each segment is a new voice; *k*-means clustering (*k*-means), as outlined above; and random clustering (Random), in which we randomly assign each voice to a cluster. For the latter two methods, both with a random component, we averaged our metrics over 50 runs. Random and Initial are here, of course, to provide baselines for judging the effectiveness of *k*-means clustering model. Finally, when using the gold-standard segmentation and *k*-means clustering, we included another oracle option (Seeded): instead of the standard *k*-means method of randomly choosing them from the available datapoints, each centroid is initialized to the longest instance of a different voice, essentially seeding each cluster.

5.3 Results

Table 3 contains the results for our first evaluation of voice clustering, the poems that were generated automatically. In all the conditions, using the gold segmentation far outstrips the other two options. The automatic segmentation is consistently better than the evenly-spaced baseline, but

Configuration	**BCubed metrics**		
	Prec.	**Rec.**	**F-score**
Initial Even	0.703	0.154	0.249
Initial Automatic	0.827	0.177	0.286
Initial Gold	1.000	0.319	0.465
Random Even	0.331	0.293	0.307
Random Automatic	0.352	0.311	0.327
Random Gold	0.436	0.430	0.436
k-means Even	0.462	0.409	0.430
k-means Automatic	0.532	0.479	0.499
k-means Gold	0.716	0.720	0.710
k-means Gold Seeded	0.869	0.848	0.855

TABLE 3 Clustering results for artificial poems

the performance is actually worse than expected. The segmentation metrics we used in the preceding section suggested that the segmentation was roughly halfway to a perfect segmentation, but the better segmentation is reflected mostly in precision and not recall; therefore clustering performance as expressed by the F-score is far less optimistic. Random clustering is clearly worse than *k*-means, but for the unreliable segmentations the harmonic mean is actually higher than the initial clustering, due to an increase in recall. The improvement due to *k*-means is sizable, and fairly consistent across the three segmentations, though better segmentations see more absolute improvement. Seeding is also quite effective, and for this relatively easy dataset we approach perfect performance under this condition.

The results for *The Waste Land* are in Table 4. Many of the basic patterns are the same, including the consistent ranking of the methods; overall, however, the clustering is far less effective. This is particularly true for the gold-standard condition, which increases only modestly between the initial and clustered state. The marked increase in recall is balanced by a major loss of precision. In fact, unlike with the artificial texts, the most promising aspect of the clustering seems to be the fairly sizable boost to the quality of clusters in automatic segmenting performance. The effect of seeding is also very consistent, nearly as effective as in the automatic case.

We also looked at the results for individual speakers in *The Waste Land*. Many of the speakers (some of whom appear only in a few lines) are very poorly distinguished, even with the gold-standard segmentation and seeding, but there are a few that cluster quite well. The best two

Configuration	BCubed metrics		
	Prec.	Rec.	F-score
Initial Even	0.792	0.069	0.128
Initial Automatic	0.798	0.084	0.152
Initial Gold	1.000	0.262	0.415
Random Even	0.243	0.146	0.183
Random Automatic	0.258	0.160	0.198
Random Gold	0.408	0.313	0.352
k-means Even	0.288	0.238	0.260
k-means Automatic	0.316	0.264	0.296
k-means Gold	0.430	0.502	0.461
k-means Gold Seeded	0.491	0.624	0.550

TABLE 4 Clustering results for *The Waste Land*

are in fact our two clearest examples of stylistic difference from Section 3.2, that is, the narrator (F-score 0.869), and the chatty woman (F-score 0.605), suggesting that our clustering behavior does correspond somewhat to a human judgment of distinctness. The former result is particularly important, from the perspective of literary analysis, since there are several passages which seem to be the main narrator (and our expert annotated them as such) but which are definitely open to interpretation. In Section 6.2, we will further explore the possibility of errors in our gold standard.

6 Profiling Voices

In this section we identify the specific stylistic characteristics of the voices of *The Waste Land* using an automatically-created lexical resource, which allows for a more human-interpretable stylistic analysis than is possible with the more eclectic range of stylistic features employed in the preceding two sections. Our goal here is to demonstrate the relevance of corpus-derived lexical style to literary corpus linguistics – where the task-driven approach we have adopted up to this point might not be particularly appealing – so as to show how this kind of information can be used to both confirm and challenge our existing interpretations of the poem.

6.1 Method

The analysis in this section relies on high-coverage stylistic lexicons, which we use to analyze stylistic differences between the various proposed characters in *The Waste Land*. The six stylistic aspects we focus on here are listed below, with the definitions adapted from earlier work (Brooke and Hirst, 2013b).

Objective Words which are emotionally distant, projecting a sense of disinterested authority. Examples include *invariable*, *finalize* and *ancillary*.

Abstract Words which refer to something that requires major psychological or cultural knowledge to grasp, which cannot purely be defined in physical terms. Examples include *sophism*, *alienation* and *implicit*.

Literary Words which one would expect to see more or less exclusively in literature; these words often feel old-fashioned or "flowery". Examples include *yonder*, *revelry* and *wanton*.

Colloquial Words which are used primarily in informal contexts, such as slang words used among friends. Examples include *booze*, *dodgy* and *crap*.

Concrete Words which primarily refer to events, objects, or properties of objects in the physical world that one would be able to see, hear, smell, or touch. Examples include *radish*, *sew* and *freeze*.

Subjective Words which are strongly emotional or reflect a personal opinion. Examples include *ugly*, *worthy* and *bastard*.

Brooke (2013) discusses in detail the rationale for these particular styles. We build our stylistic lexicons in the same way as our recent work on characterization in *To The Lighthouse* (Brooke et al., to appear), which in turn is based on work presented in various other papers (Brooke and Hirst, 2013b,a, Brooke et al., 2014, Brooke and Hirst, 2014). Our description here will therefore be fairly brief; readers looking for technical details and evaluation can see that earlier work.

Our stylistic lexicons for literature are based on the variation found across English texts in the 2010 version of Project Gutenberg.[9] The corpus consists of 24,000 texts of various genres, but to increase variation whenever possible we used heuristics to break the literature into smaller texts that distinguished narration and speech, and also the speech of different characters. Rather than relying on individual words, we segment the Project Gutenberg corpus (and *The Waste Land*) into multiword segments according to the method of Brooke et al. (2014), which allows us to capture phrases with specific stylistic connotations. Examples from *The Waste Land* include *from time to time*, *an age of*, *leave you alone*, *has no*, *made no comment*, *pick and choose*, *rose and fell*, *walking round*, *hold on tight*, *ought to be ashamed*. We do note, anecdotally, that there seem

[9] http://www.gutenberg.org/

to be fewer of these multiword chunks in *The Waste Land* than *To The Lighthouse*, perhaps because poetry tends to avoid cliché consciously.

Values for each style are assigned to every word and expression in our lexicon in two steps, both requiring a set of examples for each style. Our words come mostly from a 900-word annotation described by Brooke and Hirst (2013b), though we discarded and replaced some modern words that were inappropriate for use in older texts. In the first step, each stylistic aspect is addressed independently using the continuous lexical spectrum method of Brooke and Hirst (2014), which was shown to be superior to other corpus-based techniques for building lexicons of this kind. This model is supervised, and uses word-to-word co-occurrence probabilities as features to a ranking algorithm using support vector machines. Next, we use the initial scores as the input to the method from (Brooke and Hirst, 2013b), which improves the individual styles by considering all the styles together in a single, graphical model and updating them using label propagation.

The resulting lexicon containing all words and expressions in the multiword segmented version of *The Waste Land* was then normalized: the word with the highest score for a style received the value +1, the word with the lowest score received the value –1. To assign style scores for one or more spans, we averaged the normalized style scores for all types appearing within them. We will refer to these six numbers together as a “stylistic profile”. We used types rather than tokens so that the style scores of function words and other very common or repeated words would not unduly influence the results. Finally, to improve the readability of our results, we carried out a second normalization on the stylistic profiles by making the text as a whole the origin of our stylistic space, shifting all the results by subtracting the stylistic profile for the entire text from the stylistic profile for each character. Statistical significance was tested using an independent sample t-test.

6.2 Analysis

Cross-Character Analysis

The stylistic profiles for each character in Table 5 show that the six-style approach can capture individual characters’ peculiarities of voice, showing how Eliot distinguishes the characters in the poem. Figure 2 is a PCA projection of Table 5 into two dimensions, intended to help the reader visualize the similarities and differences among the characters. By far the most vocally distinct character in *The Waste Land* is Woman in Bar. Her style is marked by extremely high colloquial and subjective values, and extremely low values for objective and literary, which conforms with our

Character	Types	Styles					
		Lit.	**Abs.**	**Obj.**	**Col.**	**Con.**	**Sub.**
Tiresias	460	0.03	−0.04	0.09	−0.21	0.02	−0.03
Marie	132	−0.03	−0.13	−0.07	0.02	0.04	0.03
Hellfire Preacher	207	0.06	0.00	0.00	−0.14	0.07	−0.06
Chorus	105	0.03	−0.02	−0.02	0.14	0.04	−0.06
Intrepid Reporter	66	−0.03	0.28	−0.06	0.07	−0.06	0.05
Madame Sosostris	15	−0.08	0.26	−0.47	0.65	−0.14	0.16
Crazy Prufrock	399	0.00	0.07	0.01	0.01	−0.01	−0.01
Nervous One	126	0.04	0.09	−0.26	0.29	−0.07	0.07
Woman in Bar	151	−0.24	−0.01	−0.45	0.73	−0.11	0.20
The Typist	54	−0.03	0.42	0.14	−0.14	−0.02	−0.02

TABLE 5 Stylistic profiles for various characters in *The Waste Land*. Lit. = Literary, Abs. = Abstract, Obj. = Objective, Col. = Colloquial, Con. = Concrete, Sub. = Subjective.

intuitions; in fact, for most styles compared with most characters she was distinguishable at the $p < 0.001$ level. Other characters had subtler yet still distinctive stylistic profiles. Crazy Prufrock, well-educated but somewhat manic, is marked by high abstraction, high colloquialness, and high objectivity (denoting not mental stability but rather knowledge and education). The speech of Marie, an emotional, nostalgic character whose language is highly oral, is distinguished by her subjectivity, and high colloquialness. Tiresias, the narrator, is marked by his relatively low values for colloquialness and correspondingly high values of objectivity and literariness; despite preferring written forms, he is also one of the more concrete speakers, reflecting his status as a narrator. The Hellfire Preacher, who rants powerfully in the language of an Old Testament prophet, has high values for literariness, and his repeated use of physical metaphors ("There is shadow under this red rock"; "I will show you fear in a handful of dust") tends to inflate his values for concreteness.

These aggregate profiles provide grounding for differentiating characters that might otherwise seem quite similar. For instance, both Tiresias and Crazy Prufrock are well-educated, and particularly well-read in the literary classics, which they cite frequently. However, while the two are relatively similar in the literary dimension, they are strongly distinguished in colloquial ($p < 0.001$), where Prufrock's values are noticeably higher, reflecting his schizophrenic shifts across registers. In fact, Prufrock's relatively middle-of-the-road stylistic profile is due to a certain amount of canceling out of stylistic extremes, since he is far from being a stylistically inert character.

Crazy Prufrock and Nervous One, who engage in a lengthy debate in Part II of *The Waste Land*, have occasionally been read as different

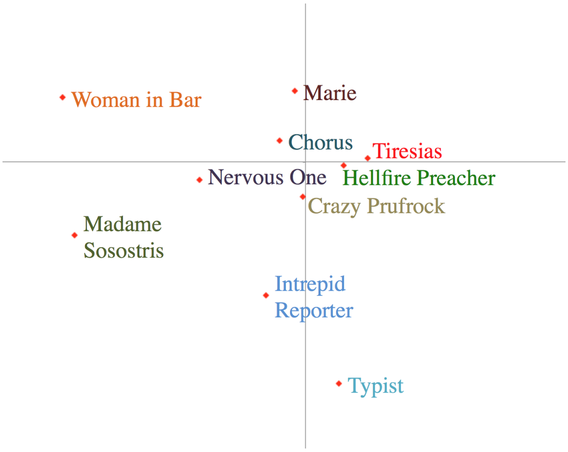

FIGURE 2 Scatterplot of the style of characters from *The Waste Land*, created by collapsing Table 5 into two dimensions using PCA

sides of a single character's split personality. Their stylistic profiles, however, are clearly distinct: the irrational and vulnerable Nervous One is much less objective, much more subjective, and much more colloquial than Prufrock (all $p < 0.001$). The poem's two most powerful prophetic voices, the Hellfire Preacher and Madame Sosostris, are likewise quite distinct: Sosostris is much less objective, much more colloquial, and much more subjective (all $p < 0.001$).

The fact that Eliot's female characters (Marie, Madame Sosostris, Nervous One, Woman in Bar, and The Typist) all have relatively high subjectivity scores may suggest a stereotyped, even sexist representation on Eliot's part. Yet in the case of Marie and Nervous One, he succeeds in distinguishing their voices in most other respects, registering statistically significant differences (all $p < 0.01$) in abstract, objective, colloquial and concrete. Indeed, across all possible pairings of characters, there were only two pairs where there was not at least one style with a statistically significant difference ($p < 0.05$): Crazy Prufrock and Chorus, which reflects that fact that these two "voices" are internally rather diverse, though for differing reasons; and Madame Sosostris and Nervous One, which may be partially attributed to similar backgrounds (both women) and registers (both oral) and lack of data for Madame Sosostris. Otherwise, based on stylistic analysis the characters seem fairly distinct, and in ways that are quite sensible given the other information we are provided about them in the poem.

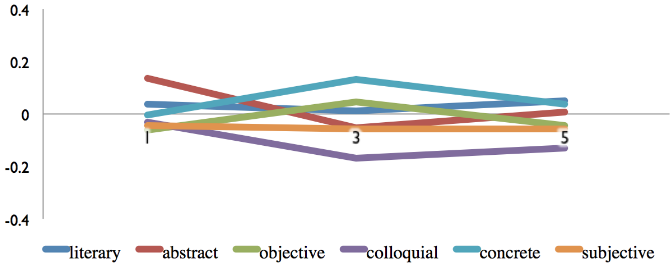

FIGURE 3 Graph of stylistic change for the Hellfire Preacher

Within-Character Analysis

Next, we looked for instances where there was significant stylistic variation across spans of the same character, to see if they might indicate a misinterpretation. Note that many spans were not long enough to lead to any conclusion, so we focused mostly on the larger spans (at least 20 word types) of fairly major characters. Figures 3- 6 each present the style scores of a particular character at different points in the text. The numbers on the x-axis are the passage numbers relative to all passages by the character in question, with only passages above that minimum size threshold included in the figures. Of those who had longer spans, some characters showed only moderate variation between individual passages of speech. For instance, the Hellfire Preacher – one of the most distinctive voices in the poem – remains consistent across all six styles in his three longest spans (see Figure 3): among his first (19-30), third (322-345), and fifth (386-392) spans, there is no statistically significant variation ($p > 0.05$). Woman in Bar was similarly consistent between her two larger spans, and we have little reason to doubt our annotation of such a distinct character.

Tiresias's function as a "narrator" of the poem – a relatively distanced, objective voice whose role it is to tell the stories of others rather than express his own feelings – results in substantial stylistic changes between his passages (see Figure 4). For example, the language in Tiresias's third passage (215-256) is significantly more colloquial ($p < 0.01$) than that of his second (77-110), and it is also markedly more subjective ($p < 0.05$). The two passages narrate quite different scenes. In the second passage, he describes a lavish scene in which a woman from a privileged social class undertakes her elaborate grooming ritual. The narrator's presentation is ironic: the description of "The chair she sat in, like a burnished throne", is deliberately overblown, and serves to show how desperately out of touch this privileged woman is with the realities of modern life

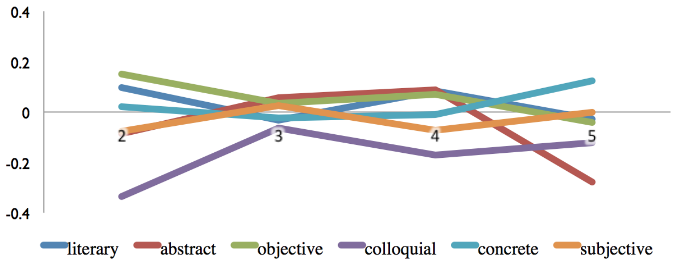

FIGURE 4 Graph of stylistic change for Tiresias

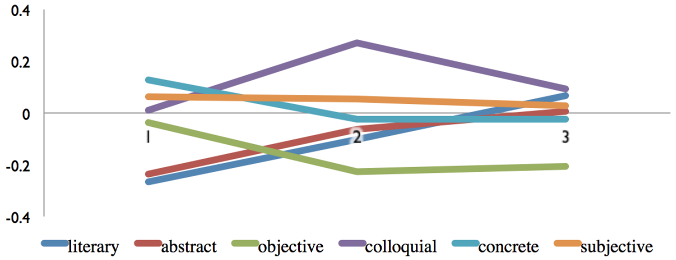

FIGURE 5 Graph of stylistic change for Marie

outside her dressing-room. The narrator spends particular time describing her painting, which presents a scene from a Greek myth that involves rape. Tiresias's third scene presents a much more direct, far less adorned, narration of an actual rape. In it, a rude, self-assured young man forces himself on the passive Typist, who is too weak and indifferent to fight him off. While these scenes respond to and mirror one another, the difference in their style reflects the honest, raw depiction of the Typist's feelings as opposed to the rich woman's closed-off emotional landscape, which is buffered on all sides by luxury items.

With Marie, too, we find some variation depending on the emotional register of the scene she is recalling or expressing (see Figure 5). In Marie's first passage (5-11), she remembers in a relatively neutral, matter-of-fact tone some scenes from her aristocratic youth: "we stopped in the colonnade, / And went on in sunlight, into the Hofgarten, / And drank coffee, and talked for an hour." In her third passage (35-41), she remembers a moment of greater emotional intensity with an old lover:

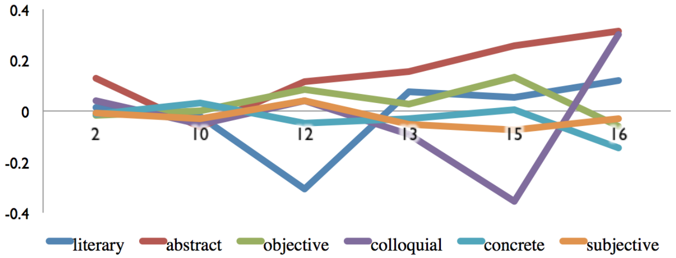

FIGURE 6 Graph of stylistic change for Crazy Prufrock

“Yet when we came back, late, from the Hyacinth garden, / Your arms full, and your hair wet, I could not / Speak, and my eyes failed, I was neither / Living nor dead, and I knew nothing, / Looking into the heart of light, the silence.” In this passage, Marie is significantly more abstract ($p < 0.05$) and literary ($p < 0.01$) than in her first, which is best explained not by positing a different character, but rather by supposing that Marie is using more poetic speech as a means of communicating the emotion of that past event.

As already noted, the character Crazy Prufrock is to some degree defined in terms of his stylistic variation (see Figure 6). His second (60-76), twelfth (207-214), and fifteenth (367-377) passages provide a representative sample of the stylistic range of this disturbed and inconsistent character. There are strong stylistic links between these passages, yet also clear differences. Passages 2 and 12 are almost identical in their values for abstract and colloquial values, but 12 has a much lower value for literary ($p < 0.01$). Passages 2 and 15 are consistent in all five styles except colloquial ($p < 0.01$). Each passage finds Prufrock in a different mental state: in his second passage, he tells a bizarre story about asking an old friend from the Crimean War whether a corpse he had planted in his garden has “begun to sprout”; in his twelfth, he provides a entirely factual account of an encounter with a merchant; in his fifteenth, he rants in a prophetic tone reminiscent of another character, the Hellfire Preacher, which seems to counter his usual colloquialness. Importantly, there is clear textual evidence linking these three passages to the same speaker: two begin with the epithet “Unreal City”, while the other ends with an evocation of the same place: “Jerusalem Athens Alexandria / Vienna London / Unreal”. This makes it unlikely that they are truly different characters, despite the stylistic differences.

Although most variation in the stylistic data for individual characters did not lead us to question our gold-standard interpretation of the poem, certain data does suggest possible errors in our interpretation. For example, one of the Prufrock inconsistencies mentioned above can be resolved if we include an extra break in one of the Prufrock passages, namely the one suggested by our automatic segmentation (as mentioned in Section 4.2), and then assign the larger part of the stanza to the Intrepid Reporter instead. Another example is the extreme stylistic discrepancy between Tiresias's fifth passage (378-385) and some of the others, particularly the second (the description of the rich woman's toilette mentioned above) (378-385), with four of the six styles showing statistically significant differences ($p < 0.05$). This passage was attributed to Tiresias because of its two references to the second passage. It begins with a description of recalling the dressing scene ("A woman drew her long black hair out tight") and it echoes the opening words "At the violet hour" in its description of "bats with baby faces in the violet light". This passage, however, also echoes many other passages from elsewhere in the poem. For instance, its description of a place where "upside down in air were towers / Tolling reminiscent bells, that kept the hours" recalls two passages in the poem, attributed to Crazy Prufrock, where towers and a clock are mentioned; and yet, stylistically the passage is a poor fit for Prufrock as well. If this passage truly merges the knowledge and perspectives of both these characters, then it should in fact be attributed to the Chorus.

7 Conclusion

Since its publication almost a century ago, *The Waste Land* has presented a unique challenge for literary scholars hoping to disentangle its cacophony of voices. The work presented here represents (to our knowledge) the first attempt to apply to this problem methods and resources derived from computational linguistics. The uniqueness of the poem represents an initial hurdle, since there is not enough information contained purely in the poem itself to built a supervised model. Much like human readers of *The Waste Land*, the approaches presented here do not rely only on the surface repetitions of words and similarly superficial features, but a deeper linguistic understanding of how words are used, one which has been derived from other sources. What we have done here, then, is not to model *The Waste Land*; on the contrary, the very specificity of *The Waste Land* requires a broad stylistic analysis. In our opinion, this is ultimately an advantage of working with individual works of literature. The inability simply to collect more training data to solve a problem like the voices of *The Waste Land* means that one must try to solve it in a way that strives for real insight into the general linguistic phenomena in question. This, in turn, results in increased applicability to related problems of the detection of stylistic inconsistency.

Sections 4-5 presented our initial, task-driven approach, involving two unsupervised modules: segmentation by looking for points of maximal stylistic change, and clustering of segments into voices. Our results show marked improvement over various baselines, but are far from exact matches with our expert gold standard. This is not surprising, since *The Waste Land* is an extremely difficult poem, even for human readers. It is also worth noting, however, that from the perspective of literary analysis there is little value in "solving" *The Waste Land*. In fact, if a completely objective gold-standard analysis of *The Waste Land* were possible, it is likely that literary scholars would never have been interested in the poem in the first place. The question of how true subjectivity can be integrated into an evaluation framework is an interesting problem (Alm, 2011), one that seems fundamental to the project of bringing techniques in computational linguistics to bear on questions of literary analysis; see our more detailed discussion of this elsewhere (Hammond et al., 2013). In any case, assuming a gold-standard analysis is useful for testing whether we are making progress towards a reasonable interpretation, but the real value in using computational methods with respect to literary analysis is in challenging existing interpretations, not simply mimicking them.

In Section 6, we addressed this rationale more explicitly, looking at lexical stylistic profiles of the voices in our gold standard to see if they truly are distinct, and whether their internal stylistic inconsistencies, when they exist, are explicable. Our methods here are more appropriate to corpus linguistics than computational linguistics, but we rely on lexical resources built using more-sophisticated statistical techniques. More generally, we want to highlight that it is not always necessary to build a full predictive model of the target phenomenon, that computational techniques can inform literary analysis simply by offering certain kinds of low-level annotation that can be then be counted and interpreted. The use of part-of-speech and semantic category tagging, for instance, is already becoming fairly common (Balossi, 2014). In certain respects, it is a much more appropriate choice with regard to our goals of being linguistically generalizable and truly cross-disciplinary. Here, the quantitative output corresponds to a real stylistic quality, and the interpretation by a literary scholar is not just a gold standard annotation: it is a key part of the result.

There are dangers in this kind of approach, however, particularly if there are too many categories resulting in spurious statistical results – which is one reason we used a small set of stylistic categories rather than the full set of features from our segmentation/cluster models – but if employed judiciously it can provide more much insight into actual stylistic variation in the poem than simply trying to optimize the parameters of a model to get the best result. With literary analysis being the only immediate “application” of work on literary texts such as *The Waste Land*, insight is more important than performance. Supporting our original annotation, we found that the voices originally posited appeared stylistically distinct and that most of the clear inconsistencies across spans of the same voice could be explained, though some of the results point in new directions for interpretation.

For the segmentation, we made the somewhat unorthodox choice of disregarding the natural formatting breaks in the poem (except as they appear in our stylistic vector), including line, clause, sentence, stanza, and part breaks, treating the text as a string of tokens. Doing this allowed us to focus on the stylistic variation in the poem. Even when a shift is clearly marked by formatting, the fact that our stylistic segmenter independently identifies it as such is encouraging, and gives weight to some of its more interesting segmentation choices, as discussed in Section 4.4.

It is another advantage of not relying on some kind of unit (beyond the token) that we could calculate our change curve using a fairly large, fixed (token) size window; this allowed for a proper stylometric comparison on either side of a point without sacrificing the ability to find any particular break simply because it appears inside some unit. From a purely task-driven perspective, however, an approach that integrated more structural information would almost certainly lead to a more accurate segmentation, though it would require a major overhaul of the method presented here as well as significant changes to other aspects of this work such as artificially-mixed poems, the baselines, etc. Another drawback to relying too much on structure is that the resulting model would almost certainly lose some of its generalizability.

There are other aspects of the segmentation and clustering methods used here that are fairly crude or arbitrary (for instance, the fixed window size, the fixed number of voices); a more-sophisticated mathematical model, e.g., a nonparametric Bayesian model, might be more appropriate. It would be ideal if the segmentation and clustering could be done within a single model, since they are clearly interrelated; our results show that it is difficult to do a proper clustering with poor segments. We would like to integrate other kinds of stylistic categories, e.g., specificity, and features, for instance bringing in extrinsic knowledge related to the style associated with larger syntactic patterns, rather than only lexical features.

One final direction for future research on *The Waste Land* is to make better use of the variety of annotations we have collected for the segmentation task.[10] In particular, collapsing the variation across annotations manually to a single gold standard, as we have done here, is somewhat unsatisfying, and trying to combine the annotations algorithmically is rife with problems (for instance, a disagreement about the location of a voice switch might result in neither break being included, or, worse, both). Unfortunately, the traditional evaluation metrics we have used here presume a single correct answer, though given their problems in other areas we should certainly consider alternatives. We do not believe there are any easy solutions which otherwise preserve the key properties of segmentation metrics, but a recently proposed metric (Fournier, 2013) based on edit-distance has addressed some of the issues with P_k and *WindowDiff* and also might be more amenable to the kinds of adaptations necessary to allow for subjective variation.

Though there are relatively few texts with the degree of stylistic mixing (and opacity with respect to the boundaries of that mixing) seen in *The Waste Land*, aspects of our work here are applicable to other modernist literature (Brooke et al., to appear) and multi-author texts (Koppel et al., 2011). We have used a modified version of our clustering model in the context of an intrinsic plagiarism task (Brooke and Hirst, 2012), though in that case we were frustrated by the use of artificial data which did not reflect the real-world task. Although stylistic inconsistency (intentional or otherwise) is a common phenomenon in language, there is very little annotated data, and so data collection should be a top priority going forward. In the meantime, *The Waste Land* should serve as an intriguing test case for future computational work in this area.

Acknowledgments

This work was financially supported in part by the Natural Sciences and Engineering Research Council of Canada. The second author was supported by the Social Sciences and Humanities Research Council of Canada.

[10] All the annotations, including those by the students and experts as well as the final gold standard are available at http://www.cs.toronto.edu/~jbrooke/wasteland_annotations.zip.

References

Alm, Cecilia Ovesdotter. 2011. Subjective Natural Language Problems: Motivations, Applications, Characterizations, and Implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112.

Amigó, Enrique, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval* 12(4):461–486.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta. ISBN 2-9517408-6-7.

Bagga, Amit and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the Vector Space Model. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL-COLING '98)*, pages 79–85. Montreal, Quebec, Canada.

Balossi, Giuseppina. 2014. *A Corpus Linguistic Approach to Literary Language and Characterization: Virginia Woolf's The Waves*. Philadelphia: John Benjamins.

Beatie, Bruce A. 1967. Computer study of medieval German poetry: A conference report. *Computers and the Humanities* 2(2):65–70.

Bedient, Calvin. 1986. *He Do the Police in Different Voices: The Waste Land and its protagonist*. Chicago: University of Chicago Press.

Beeferman, Doug, Adam Berger, and John Lafferty. 1999. Statistical Models for Text Segmentation. *Machine Learning* 34:177–201.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge University Press.

Blei, David M. and Pedro J. Moreno. 2001. Topic segmentation with an aspect hidden Markov model. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 343–348. ISBN 1-58113-331-6.

Brants, Thorsten and Alex Franz. 2006. *Web 1T 5-gram Corpus Version 1.1*. Google Inc.

Brooke, Julian. 2013. *Computational Approaches to Style and the Lexicon*. Ph.D. thesis, University of Toronto, Toronto, ON, Canada. http://www.cs.toronto.edu/pub/gh/Brooke-PhD-thesis.pdf.

Brooke, Julian, Adam Hammond, and Graeme Hirst. 2012. Unsupervised Stylistic Segmentation of Poetry with Change Curves and Extrinsic Features. In *Proceedings of the 1st Workshop on Computational Literature for Literature (CLFL '12)*. Montreal.

Brooke, Julian, Adam Hammond, and Graeme Hirst. To appear. Using Models of Lexical Style to Quantify Free Indirect Discourse in Modernist Fiction. Digital Scholarship in the Humanities.

Brooke, Julian and Graeme Hirst. 2012. Paragraph clustering for intrinsic plagiarism detection using a stylistic vector-space model with extrinsic features. In *Notebook for PAN 2012 Lab at CLEF '12*. Rome.

Brooke, Julian and Graeme Hirst. 2013a. A multi-dimensional Bayesian approach to lexical style. In *Proceedings of the 13th Annual Conference of the North American Chapter of the Association for Computational Lingusitics*.

Brooke, Julian and Graeme Hirst. 2013b. Hybrid Models for Lexical Acquisition of Correlated Styles. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP '13)*.

Brooke, Julian and Graeme Hirst. 2014. Supervised Ranking of Co-Occurrence Profiles for Acquisition of Continuous Lexical Attributes. In *Proceedings of The 25th International Conference on Computational Linguistics (COLING 2014)*.

Brooke, Julian, Graeme Hirst, and Adam Hammond. 2013. Clustering voices in *The Waste Land*. In *Proceedings of the 2nd Workshop on Computational Literature for Literature (CLFL '13)*. Atlanta.

Brooke, Julian, Vivian Tsang, Graeme Hirst, and Fraser Shein. 2014. Unsupervised Multiword Segmentation of Large Corpora Using Prediction-Driven Decomposition of n-Grams. In *Proceedings of The 25th International Conference on Computational Linguistics (COLING 2014)*.

Brooke, Julian, Tong Wang, and Graeme Hirst. 2010. Automatic Acquisition of Lexical Formality. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*. Beijing.

Burnard, Lou. 2000. User reference guide for British National Corpus. Tech. rep., Oxford University.

Burrows, John F. 1987. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.

Burton, Kevin, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r Dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM '09)*. San Jose, CA.

Cooper, John Xiros. 1987. *T.S. Eliot and the politics of voice: The argument of The Waste Land*. Ann Arbor, Mich.: UMI Research Press.

Cox, David R. and Peter A.W. Lewis. 1966. *The Statistical Analysis of Series of Events*. Monographs on Statistics and Applied Probability. Chapman and Hall. ISBN 9780412218002.

Culpeper, Jonathan. 2009. Keyness: Words, Parts-Of-Speech and Semantic Categories in the Character-Talk of Shakespeare's *Romeo And Juliet*. *International Journal of Corpus Linguistics* 14(1):29–59.

Dale, Edgar and Jeanne Chall. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Cambridge, MA: Brookline Books.

DeForest, Mary Margolies and Eric Johnson. 2000. Computing Latinate word usage in Jane Austen's novels. *Computers & Texts* 18/19:24–25.

Duggan, Joseph J. 1973. *The Song of Roland: Formulaic style and poetic craft*. University of California Press.

Eder, Maciej. 2015. Rolling stylometry. *Digital Scholarship in the Humanities* .

Eisenstein, Jacob and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*.

Emigh, William and Susan C. Herring. 2005. Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS '05)*. Washington, DC.

Fournier, Chris. 2013. Evaluating Text Segmentation using Boundary Edit Distance. In *Proceedings of 51st Annual Meeting of the Association for Computational Linguistics (ACL '13)*.

Galley, Michel, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL '03)*, pages 562–569. Sapporo, Japan.

Graham, Neil, Graeme Hirst, and Bhaskara Marthi. 2005. Segmenting documents by stylistic character. *Natural Language Engineering* 11(4):397–415.

Guthrie, David. 2008. *Unsupervised Detection of Anomalous Text*. Ph.D. thesis, University of Sheffield.

Hammond, Adam, Julian Brooke, and Graeme Hirst. 2013. A Tale of Two Cultures: Bringing Literary Analysis and Computational Linguistics Together. In *Proceedings of the 2nd Workshop on Computational Literature for Literature (CLFL '13)*.

Hearst, Marti A. 1994. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL '94)*, pages 9–16.

Heylighen, Francis and Jean-Marc Dewaele. 2002. Variation in the Contextuality of Language: An Empirical Measure. *Foundations of Science* 7(3):293–340.

Kao, Justine and Dan Jurafsky. 2012. A Computational Analysis of Style, Sentiment, and Imagery in Contemporary Poetry. In *Proceedings of the 1st Workshop on Computational Linguistics for Literature (CLFL '12)*. Montreal.

Kazantseva, Anna and Stan Szpakowicz. 2014. Hierarchical Topical Segmentation with Affinity Propagation. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*.

Kestemont, Mike, Kim Luyckx, and Walter Daelemans. 2011. Intrinsic plagiarism detection using character trigram distance scores. In *Proceedings of the PAN 2011 Lab: Uncovering Plagiarism, Authorship, and Social Software Misuse*.

Koppel, Moshe, Navot Akiva, Idan Dershowitz, and Nachum Dershowitz. 2011. Unsupervised decomposition of a document into authorial components. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL '11)*. Portland, Oregon.

Landauer, Thomas K. and Susan Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review* 104:211–240.

MacQueen, J. B. 1967. Some Methods for Classification and Analysis of MultiVariate Observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.

Malioutov, Igor and Regina Barzilay. 2006. Minimum cut model for spoken lecture segmentation. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL '06)*, pages 25–32. Sydney, Australia.

McKenna, C. W. F. and A. Antonia. 2001. The statistical analysis of style: Reflections on form, meaning, and ideology in the 'Nausicaa' episode of *Ulysses*. *Literary and Linguistic Computing* 16(4):353–373.

Oberreuter, Gabriel, Gaston L'Huillier, Sebastián A. Ríos, and Juan D. Velásquez. 2011. Approaches for intrinsic and external plagiarism detection. In *Proceedings of the PAN 2011 Lab: Uncovering Plagiarism, Authorship, and Social Software Misuse*.

Pevzner, Lev and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics* 28(1):19–36.

Schmid, Helmut. 1995. Improvements In Part-of-Speech Tagging With an Application To German. In *Proceedings of the ACL SIGDAT Workshop*, pages 47–50.

Sigg, Eric. 1994. Eliot as a Product of America. In A. D. Moody, ed., *The Cambridge Companion to T. S. Eliot*, pages 14–30. Cambridge: Cambridge University Press.

Simonton, Dean Keith. 1990. Lexical choices and aesthetic success: A computer content analysis of 154 Shakespeare sonnets. *Computers and the Humanities* 24(4):251–264.

Stamatatos, Efstathios. 2009. Intrinsic plagiarism detection using character *n*-gram profiles. In *Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and, Social Software Misuse (PAN-09)*, pages 38–46. CEUR Workshop Proceedings, volume 502.

Stein, Benno, Nedim Lipka, and Peter Prettenhofer. 2011. Intrinsic plagiarism analysis. *Language Resources and Evaluation* 45(1):63–82.

Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilivie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics* 37(2):267–307.

Touch Press LLP. 2011. *The Waste Land* app. http://itunes.apple.com/ca/app/the-waste-land/id427434046?mt=8.

Utiyama, Masao and Hitoshi Isahara. 2001. A statistical model for domain-independent text segmentation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL '01)*, pages 499–506. Toulouse, France.

Voigt, Rob and Dan Jurafsky. 2013. Tradition and Modernity in 20th Century Chinese Poetry. In *Proceedings of the 2nd Workshop on Computational Linguistics for Literature (CLFL '13)*. Atlanta.

Wallace, Byron C. 2012. Multiple Narrative Disentanglement: Unraveling *I*nfinite Jest. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT '12)*.

Wiebe, Janyce M. 1994. Tracking point of view in narrative. *Computational Linguistics* 20(2):233–287.