Unsupervised Stylistic Segmentation of Poetry with Change Curves and Extrinsic Features

Julian Brooke
Dept of Computer Science
University of Toronto
jbrooke@cs.toronto.edu

Adam Hammond
Dept of English
University of Toronto
adam.hammond@utoronto.ca

Graeme Hirst
Dept of Computer Science
University of Toronto
gh@cs.toronto.edu

Abstract

The identification of stylistic inconsistency is a challenging task relevant to a number of genres, including literature. In this work, we carry out stylistic segmentation of a well-known poem, *The Waste Land* by T.S. Eliot, which is traditionally analyzed in terms of numerous voices which appear throughout the text. Our method, adapted from work in topic segmentation and plagiarism detection, predicts breaks based on a curve of stylistic change which combines information from a diverse set of features, most notably co-occurrence in larger corpora via reduced-dimensionality vectors. We show that this extrinsic information is more useful than (within-text) distributional features. We achieve well above baseline performance on both artificial mixed-style texts and *The Waste Land* itself.

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.

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. Though he worked in a variety of forms — he was a celebrated critic as well as a dramatist, receiving a Tony Award in 1950 — he is best remembered today for his poems, of which *The Waste Land* (1922) is among the most famous. The poem deals with themes of spiritual death and rebirth. It is notable for its disjunctive structure, its syncopated rhythms, its wide range of literary allusions, and its incorporation of numerous other languages. 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). Eliot himself, in his notes to *The Waste Land*, supports the latter view by referring to "characters" and "personage[s]" in the poem.

One of the poem's most distinctive voices is that of the woman who speaks at the end of its second section:

Workshop on Computational Linguistics for Literature, pages 26–35,
Montréal, Canada, June 8, 2012.

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]

Her chatty tone and colloquial grammar and lexis distinguish her voice from many others in the poem, such as the formal and traditionally poetic voice of a narrator 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]

While the stylistic contrasts between these and other voices are apparent to many readers, Eliot does not explicitly mark the transitions between them. The goal of the present work is to investigate whether computational stylistic analysis can identify the transition between one voice and the next.

Our unsupervised approach, informed by research in topic segmentation (Hearst, 1994) and intrinsic plagiarism detection (Stamatatos, 2009), 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 artificial (mixed) poems, but the our main evaluation is on *The Waste Land* itself. We believe that even when our segmentation disagrees with expert human judgment, it has the potential to inform future study of this literary work.

2 Related work

Poetry has been the subject of extensive computational analysis since the early days of literary and linguistic computing (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 with 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 Ronald* manifests primarily an oral or a written style. Also related to our work, although it concerned a novel rather than a poem, is that of 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.

More general work on identifying stylistic inconsistency includes that of Graham et al. (2005), who built artificial examples of style shift by concatenating Usenet postings by different authors. 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 for 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.

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 that 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 that compares, using a special distance function, a character trigram feature vector at various steps throughout the text, creating a style change function whose maxima indicate points of interest (potential plagarism).

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.

3 Stylistic change curves

Many popular text segmentation methods depend crucially on a reliable textual unit (often a sentence) which can be reliably classified or compared to others. But, for our purposes here, a sentence is both too small a unit — our stylistic metrics will be more accurate over larger spans — and not small enough — we do not want to limit our breaks to sentence boundaries. 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 the TextTiling technique of Hearst (1994) and the style change function of Stamatatos (2009), but our model is generalized so that it applies to any numeric metric (feature) 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 at point c_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 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 distance, and city-block distance. In his study, Guthrie (2008) found best results with city-block distance, and that is what we will primarily use here. 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. However,

stylistic change curves 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, however, 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 $g_i > g_j$. 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 Features

The set of features we explore for this task falls 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 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.

4.1 Surface Features

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 for Guthrie (2008).

Syllable count Syllable count is 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, semicolons can be 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 dynamically including 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 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.

4.2 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). 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 lexicons provide a measure of the degree to which a word is positive or negative. Instead of summing the sentiment scores, we sum their absolute values, to get a measure of how extreme (subjective) the span is.

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 highly diverse register/genre corpora, 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 provided the best results in a near-synonym evaluation. Early work by Biber (1988) in the Brown Corpus using a related technique (factor analysis) resulted in discovery of several identifiable dimensions of register. Here, we investigate using these LSA-derived vectors directly, 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 is 100 million tokens. 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.

5 Evaluation method

5.1 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. (1997), uses a probe window equal 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 (according to the reference) 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 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). 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. 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 fairly serious shortcoming with 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 under 0.3 (a very competitive result both here and in topic segmentation), if the variation of the true segment size is reasonably high. 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 therefore use a third 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 also flawed, because it can be trivially made 0 (the best score) by guessing a break everywhere. However, the relative motion of the two kinds of metric provides insight into whether we are simply moving along a precision/recall curve, or actually improving overall segmentation.

5.2 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 use the 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 use the average score over 50 runs.

6 Experiments

6.1 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 from different authors, as others have done with prose.

6.1.1 Setup

Our set of twelve poems 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 *The Waste Land* in particular. The titles were removed, and each poem was tagged by an automatic PoS tagger (Schmid, 1995). Koppel et al. 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 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 which minimize the sum of the three metrics.

6.1.2 Results

The Feature Sets section of Table 1 gives the individual feature results for segmentation of 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 did more modestly, though the extremeness of polarity was useful when paired with the coverage of SentiWordNet. Among the larger feature sets, the GI was the least useful, though more effective than any of the

Table 1: Segmentation accuracy in artificial poems

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
Dynamic (*tf-cl*, 1000)	0.344	0.333	0.199
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
GI (*tf-cl*, 100)	0.388	0.376	0.163
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

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 4 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.

6.2 *The Waste Land*

6.2.1 Setup

In order to evaluate our method on *The Waste Land*, we first created a gold standard voice switch segmentation. 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 based on their own intuitions, and we created a combined student version based on majority judgment. Second, our English literature expert listened to the 6 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; these were combined to create a reader version. Finally, our expert amalgamated these two versions and incorporated insights from independent literary analysis to create a final gold standard.

We created two versions of the poem for evaluation: for 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 involves removing all segments (and the corresponding voice switches, when appropriate) which are 20 or fewer tokens in length and/or which are in a language other than English, which 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).

6.2.2 Results

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 GI is too unreliable 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 with real switches and found that there were several instances (corresponding to very clear voices switches in the text) which were nearly perfect. 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 windows never consisted of a pure style).

Table 2: Segmentation accuracy in *The Waste Land*

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**

7 Conclusion

In this paper we have presented a system for automatically segmenting stylistically inconsistent text and applied it to *The Waste Land*, a well-known poem in which stylistic variation, in the form of different 'voices', provides an interesting challenge to both human and computer readers. Our unsupervised model is based on a stylistic change curve derived from feature vectors. Perhaps our most interesting result is the usefulness of low-dimension LSA vectors over surface features such as words and trigram characters as well as other extrinsic features such as the GI dictionary. In both *The Waste Land* and our development set of artificially combined poems, our method performs well above baseline. Our system could probably benefit from the inclusion of machine learning, but our main interest going forward is the inclusion of additional features — in particular, poetry-specific elements such as alliteration and other more complex lexicogrammatical features.

Acknowledgments

This work was financially supported by the Natural Sciences and Engineering Research Council of Canada.

References

Stefano Baccianella, 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, May.

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

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

Doug Beeferman, Adam Berger, and John Lafferty. 1997. Text segmentation using exponential models. In *In Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP '97)*, pages 35–46.

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

David M. Blei 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.

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

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

Lou Burnard. 2000. User reference guide for British National Corpus. Technical report, Oxford University.

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

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

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

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

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

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08*, EMNLP '08, pages 334–343.

William Emigh 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.

Michel Galley, 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)*, ACL '03, pages 562–569.

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

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

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

Francis Heylighen and Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science*, 7(3):293–340.

Mike Kestemont, 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*.

Moshe Koppel, 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)*.

Thomas K. Landauer 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.

Igor Malioutov 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.

C. W. F. McKenna 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.

Gabriel Oberreuter, 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*.

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

Helmut Schmid. 1995. Improvements in part-of-speech tagging with an application to German. In *Proceedings of the ACL SIGDAT Workshop*, pages 47–50.

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

Efstathios Stamatatos. 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.

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

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

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manifred 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 .

Masao Utiyama 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.