IN SEARCH FOR VOLTA: STATISTICAL ANALYSIS OF WORD PATTERNS IN SHAKESPEARE'S SONNETS

Oskar Kohonen[1]*, Sakari Katajamäki*[2] *and Timo Honkela*[1]

[1] Neural Networks Research Centre, Helsinki University of Technology,
P.O.Box 5400, FIN-02015 Espoo, FINLAND, {oskar.kohonen, timo.honkela}@hut.fi
[2] Department of Comparative Literature, Institute for Art Research, University of Helsinki
FIN-00014, FINLAND, sakari.katajamaki@helsinki.fi

ABSTRACT

The sonnet is one of the most canonical modes of poetry in Western literature. The English or Shakespearean sonnet falls in to three quatrains in iambic pentameter, with a turn at the end of the line 12 and a concluding couplet often of a summary or epigrammatic character. The turn normally is both semantic and stylistic for the rhyme scheme *abab cdcd efef* of the first part of the poem changes to a form *gg* in the closing couplet. We analyze if the semantic turn, or volta, can be found with statistical analysis of word distributions, using the Self-Organizing Map for exploration and visualization. The Self-Organizing Map is a neural network architecture based on unsupervised learning. We conclude that our methods can be useful for finding semantic and stylistic turns that can then be studied in detail using other methods. We propose extensions to our methods for other literary analysis.

1. INTRODUCTION

In this article we present a way to analyze poems, in particular, Shakespeare's sonnets. Our specific focus is to see what kind information we can find on the "semantic turn" in a sonnet. This is a topic that is related both to the structure of a poem and the meaning of the words used. Our aim is not to present an analysis model that would cover all relevant aspects but to outline one particular approach that can be later extended to cover other points of view. In the following, we introduce the concept of a sonnet, and in the subsequent chapters we describe the method that we have used and the results obtained.

The sonnet is one of the most canonical modes of poetry in Western literature. It originated in Sicily in the 13th century, since being introduced to all Western countries. What is characteristic for all sonnets – with a few exceptions – is that they consist of fourteen lines that follow certain rhyme scheme and that they have in certain point a 'turn', a shift in direction or tone, often further emphasised by a stanza-break.

For instance, in the original Italian sonnet form, Petrarchan sonnet, the sonnet's rhyme scheme divides the poem's 14 lines into two parts, an octave (an eight-line unit of the two first stanzas, rhymed abba abba) and a sestet (a six-line unit of the last two stanzas, typically rhymed *cdc dcd*). In the Italian sonnet form, the turning point, the so-called volta, is between the octave and the sestet where the rhyme scheme changes.

The English or Shakespearean sonnet falls in to three quatrains (four-line units) in iambic pentameter (five feet with an unstressed followed by a stressed syllable, as in 'deceive'), with a turn at the end of the line 12 and a concluding couplet often of a summary or epigrammatic character. The turn normally is both semantic and stylistic for the rhyme scheme *abab cdcd efef* of the first part of the poem changes to a form *gg* in the closing couplet.

In part, the reason why the genre of the sonnet has remained so popular even 800 years is the sonnet's strictly limited structure based on formal and semantic rules. Besides, literary theorists and linguists have found the sonnet's clear basic structure a good corpus for different kinds of theoretical inquiries. Roman Jakobson and Claude Lévi-Strauss (1962) have used Baudelaire's sonnet "Les Chats" in improving structuralist methods, Alastair Fowler (1982, 2000) illustrates his theories on the transformation of literary genres by using sonnets, and Jerry R. Hobbs (1990) has developed cognitive analysis using sonnets as a corpus.

Previously, statistical analysis of Shakespeare's language in his Sonnets and other works have been used, for instance, to distinguish works really written by Shakespeare from counterfeits, anonymously authored works, wrongly presented as Shakespeare's (cf. Vickers 2002).

Shakespeare's sonnets are an excellent corpus for the present kind of methodological experiment, in which we try to explore the relationships between lexical changes and semantic or stylistic turns in the sonnets. Firstly, the sonnet form, by definition, provides us a clear hypothesis of the "right" place of the turning point, if it is possible to find using a lexico-statistical method. Secondly, Shakespeare's sonnets, as a collection of love poems, are thematically similar, which may help the statistical analysis. Thirdly, the reasonable amount of the analyzed sonnets, 152 sonnets, makes it possible to compare the results of the statistical analysis with the analyses made by hand based on human understanding and intuition.

2. DATA AND METHODS

In our experiment the data consisted of a collection of 154 Shakespeare sonnets. Out of these sonnets two (numers 99 and 126) were not written in the regular sonnet form, but consisted of 12 and 15 lines instead of 14. These two sonnets were removed from the analysis. The remaining sonnets were tokenized into words simply by lowercasing everything, removing every special sign, except for the apostrophe, and finally splitting at every whitespace. What remains is a stream of words with morphological markers intact (for instance "beauty's"). In total the stream consisted of 17000 words in the "running text" (tokens), and 3300 separate words (types). For the self-organizing map analysis we selected those words which appeared ten or more times in the text, 228 words in total.

Our goal was to evaluate if an analysis of word distributions could find turning points (e.g. volta) in the sonnets. These distributions can be represented by a vector for each word. The vectors can then be analyzed using the self-organizing map. The self-organizing map is a neural network architecture based on unsupervised learning (Kohonen 2001).

Initially we experimented with a fourteen dimensional vector, each dimension corresponding to a line in the sonnet. For example if the word "for" occurs twice on the first line, three times on the fifth, four times in the twelfth, once in the fourteenth and none otherwise, the word "for" will be represented by a vector:

$v = [2, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 4, 0, 1]$.

This approach did not result in very good visualizations, because the data was very sparse. Therefore we decided, following the sonnet structure, to use only a four-dimensional vector to represent the words. The first dimension of the vector is the sum of occurrences of the word in the first four lines, the second dimension is the same for the following four lines, as is also the third dimension. The fourth dimension corresponds only to the last two lines, to compensate for this we multiply those occurrences by two. The representation for "for", above, then becomes:

$v = [2+0+0+0, 3+0+0+0, 0+0+0+4, 2*(0+1)]$
$= [2, 3, 4, 2]$

3. RESULTS

Using the four-dimensional vector representation we constructed a self-organizing map of the word distributions. The self-organizing map creates a topology preserving mapping from the four-dimensional vector space to a two-dimensional map, which is useful for visualization and exploratory analysis. Those words that are similar in the four-dimensional space are close to each other on the map.

We show four different images of the same map. The different figures are different views of the same map and should be viewed at the same time. The image in Fig. 1 shows specific words on the map.

Because of the word encoding the order on the map shown in Fig. 1 does not straightforwardly reflect, e.g., syntactical or semantic categories (like in the analysis based

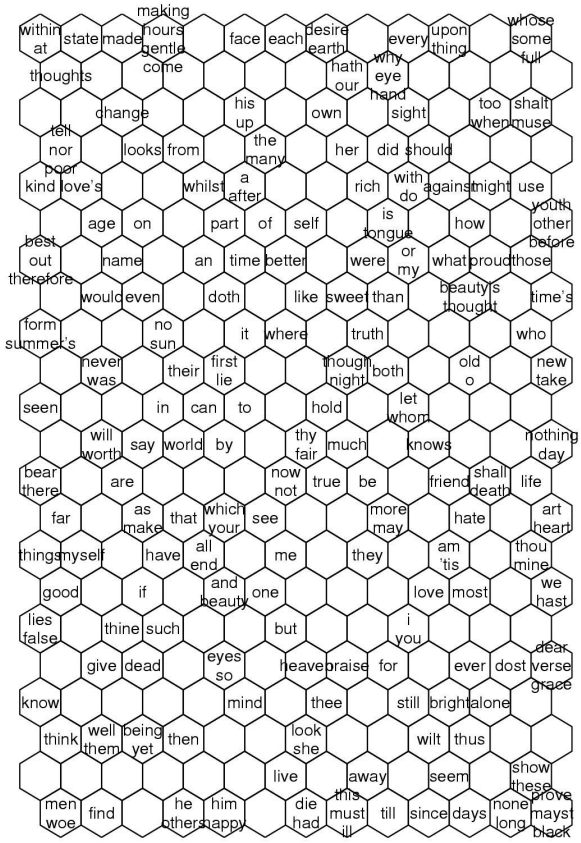

Figure 1. The 228 most frequent words in the corpus on a self-organizing map. The order is based on the four-dimensional distribution of each word in the quatrains of the sonnets. Because of the word encoding, the order on the map does not straightforwardly reflect, e.g., syntactical or semantic categories, but rather reflects the stylistic choices of the author. Specifically, the position on the map is based on where in the sonnets the words typically appear.

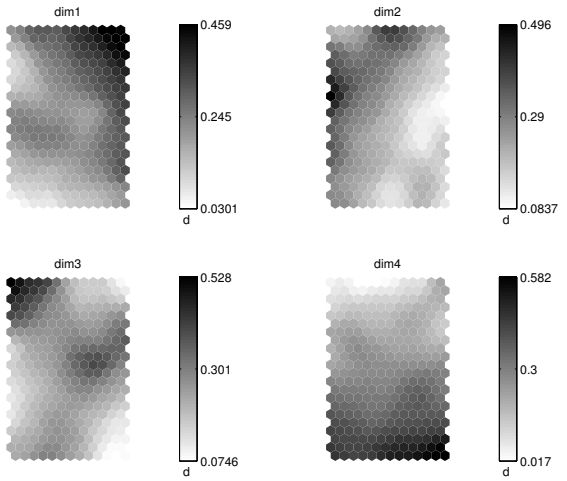

Figure 2. The values of the different dimensions on the map, darker color means bigger value. Notice how the fourth dimension is high towards the bottom of the map while the other dimensions are lower in that area.

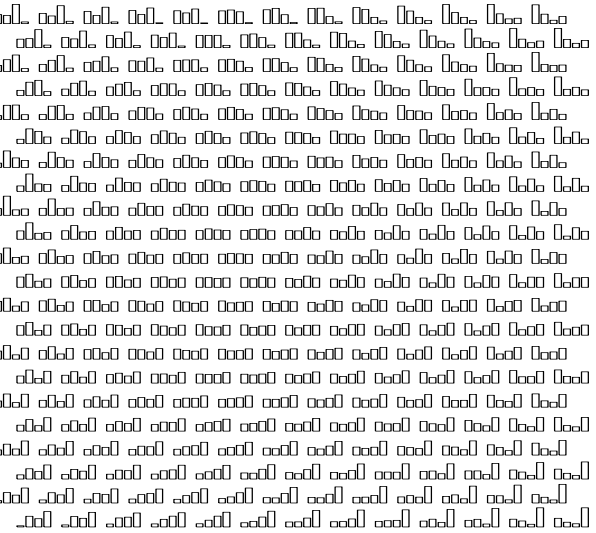

Figure 3. The distributions of the map elements shown as bar diagrams. The distributions in a certain cell correspond to the distributions of the words in the same cell on the map in figure 1. In one cell, the leftmost bar indicates the number of occurrences in the first quatrain and so on. The middle region tends towards flat distributions and the corners to distributions tilted in different directions.

on words in their contexts, cf. (Honkela et al. 1995, Lagus et al. 2002). This map rather reflects the stylistic choices of the author. More specifically, the position on the map is based on where in the sonnets the words typically appear.

Fig. 2 shows how the values of the different dimensions in the vectors vary across the map. Darker color means higher value of that dimension. As can be seen the different four-dimensions have high values on different parts of the map, for instance dimension one has a high value in the upper right corner and a small one in the lower left corner, and dimension three is almost a mirror image of this. The first image shows that the words "prove" and "black" are in the lower right corner of the map. In combination with the second we can conclude that these words are frequent in the couplet after the volta, and not as frequent in other positions. The specific distributions for these words are: $v_{prove} = [3, 2, 2, 10]$ and $v_{black} = [4, 1, 3, 10]$. (note that the last dimension is the occurrences of the word in the couplet multiplied by two, whereas the others are the number of occurrences).

Similarly the words "whose" and "full" appear frequently in the first quatrain, but seldom elsewhere. Their distributions are: $v_{whose} = [8, 6, 2, 6]$ and $v_{full} = [6, 4, 1, 4]$.

As can be seen, from the distributions of "whose" and "full", there is a tendency towards the first dimension, but it does not mean that these words only exist in the first quatrain. A visualization of how the distributions vary over the map can be seen in Fig.3.

That the high values of the dimensions have occupied different regions of the map, means that the dimensions are quantifiably different, even though this is not easily seen from the distributions shown above. In this specific case it means that the vocabularies in the four different parts of the sonnet are different from each other. The self-organizing map manages to find differentiating structure from the tendencies of the distributions to shift in different directions. Especially interesting, considering our goal to find the turning point, is the fact that the fourth dimension corresponding to the lines after the volta are quite different from the others. Whereas the others tend towards the top of the map, large values in the fourth dimension occupy the region at the bottom, apart from the others.

If cells are adjacent on the map, it means that the data they represent are similar. However, since the transformation from the four-dimensional space to the two-dimensional map is non-linear, the distances between adjacent cells on the map vary in the original four-dimensional space. A U-matrix (Fig. 4 visualizes the distances between adjacent cells on the map in the original four-dimensional space.

It shows that the upper left corner, a part of the left side and the lower right corner are far away from the surrounding nodes. This can be seen from the darker areas in the U-matrix. From the U-matrix we can also tell that even if the fourth dimension is different from the others, the boundary is not very sharp. If the boundary would be sharp, there would also be a darker "mountain" area between the bottom part of the map and the part above it. As can be seen, there is no such mountain. Only the lower right corner is separated by a darker part in the U-matrix. Otherwise the U-matrix is flat, which suggests a gradual increase of the fourth dimension, instead of a sharp boundary. This is disappointing from a classification perspective, as it means that single words do not predict the position after the volta very well. In this way the self-organizing map shows, that even though there is structure in the data set, the structure is not very crisp, and might not be good for classification.

The analysis with the self-organizing map suggests

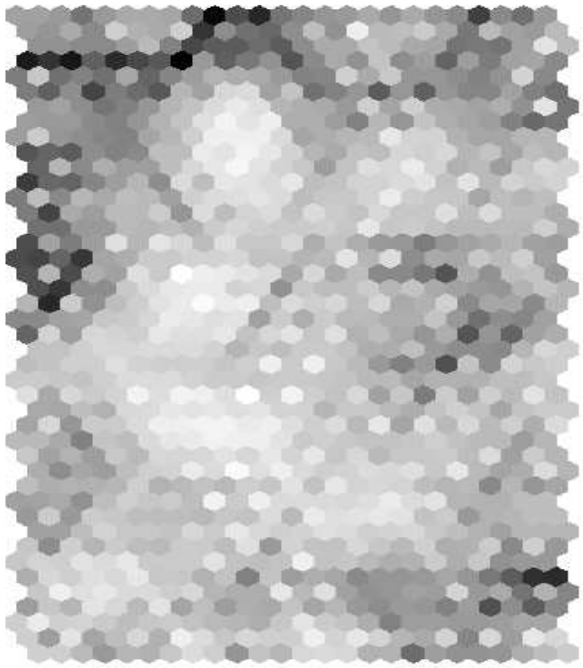

Figure 4. The distances between the different cells on the map. Here darker means longer distance. The shade of gray indicates clustering on map and it can metaphorically be thought of as a landscape in which the darker regions are kind of mountain ranges that separate valleys from each other.

that we could find the volta in Shakespeare's sonnets, with some accuracy, using only the frequencies of certain words in the surrounding lines. The occurrence of the words that appear in the lower right corner on the map predict that we have passed the volta. Similarly words in the upper left corner predict that we are in the third quatrain. The words in the middle of the map have flat distributions, that is they do not predict any particular part of the sonnet. The analysis also suggests that a classification task of this kind would not be very easy, and if such a task is attempted, it is recommended that the word distributions are combined into more predictive features or that some other more predictive features are extracted. Moreover, for prediction other methods based, e.g., Bayesian principles would also be useful.

4. CONCLUSIONS

A stylistic, semantic or psychological turning point in a literary text does not necessarily mean a change of the word frequencies at the same place. However, lexical analysis is one way of searching for different kinds of turning points in literature.

For a literary analysis, the method presented in this paper, is probably most useful in the analysis of large data sets. The advantage of the self-organizing maps analysis, is that it can be performed in an exploratory fashion and provides a fairly succint visualization of the structure of the word distributions. Therefore, it can be used as a preliminary survey, as in this paper, that provides us with a method for making hypotheses that can be further studied using other methods. In analysing volta-like turning points, it is useful that the corpus follows certain formal homogeneity (for instance, such genres as haiku, limerick or certain folk song genres). However, it would be possible to expand the analysis to turning points in prose by using a suitable windowing scheme (for instance, one page or one chapter corresponds to a fixed number of dimensions in the vectorial representation). Such extensions might require certain restrictions on the vocabulary included in the analysis or additional preprocessing steps.

For poems with a strict form, an interesting alternative would be using smaller units than words as a basis. For a regular rhyming scheme, one could use phonemes, or even approximate them with letters in some cases. This could yield interesting results on whether other regular structure exists in the poems apart from the known rhyming scheme. Moreover, one could also consider analysis in the level of syllables or morphemes. It is even possible to segment words into morpheme-like units in an unsupervised manner (Creutz and Lagus 2002).

Expanding the analysis for rhyming corpora without formal structure, such as contemporary music lyrics, would be challenging. This would require another approach to structure, than the rigid line based scheme we have used.

References

Biggs, Henry P. (1996), *A Statistical Analysis of the Metrics of the Classic French Decasyllable and Classic Alexandrine*. PhD Thesis, Los Angeles, CA: UCLA.

Creutz, Mathias and Lagus, Krista (2002), Unsupervised discovery of morphemes. In: *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, pp. 21-30.

Fowler, Alastair (1982), *Kinds of Literature*. Oxford: Oxford University Press.

Fowler, Alastair (2000), *"Transformations of Genre"*. In: Duff, David (ed.), Modern Genre Theory. New York: Longman, pp. 232-249.

Hobbs, Jerry R. (1990), *Literature and Cognition*. Stanford, CA: CSLI.

Honkela, Timo, Pulkki, Ville, and Kohonen, Teuvo (1995), Contextual relations of words in Grimm tales analyzed by self-organizing map. In: *Proceedings of ICANN-95, International Conference on Artificial Neural Networks*, F. Fogelman-Soulié and P. Gallinari (eds), vol. 2, EC2 et Cie, Paris, pp. 3-7.

Jakobson, Roman and Lévi-Strauss, Claude (1962), *"Les Chats' de Charles Baudelaire"*. L'Homme 2.

Kohonen, Teuvo (2001), *Self-Organizing Maps*. Berlin, Heidelberg, New York: Springer.

Lagus, Krista, Airola, Anu, and Creutz, Mathias (2002), Data analysis of conceptual similarities of Finnish verbs. In: *Proceedings of CogSci 2002, the 24th annual meeting of the Cognitive Science Society*.

Vickers, Brian (2002), *Counterfeiting Shakespeare*. Cambridge University Press.