

Exploring Shakespeare's Sonnets with SPARSAR

Rodolfo Delmonte

Department of Language Studies & Department of Computer Science, Ca' Foscari University, Italy

Copyright © 2016 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract Shakespeare's Sonnets have been studied by literary critics for centuries after their publication. However, only recently studies made on the basis of computational analyses and quantitative evaluations have started to appear and they are not many. In our exploration of the Sonnets we have used the output of SPARSAR which allows a full-fledged linguistic analysis which is structured at three macro levels, a *Phonetic Relational Level* where phonetic and phonological features are highlighted; a *Poetic Relational Level* that accounts for a poetic devices, i.e. rhyming and metrical structure; and a *Syntactic-Semantic Relational Level* that shows semantic and pragmatic relations in the poem. In a previous paper we discussed how colours may be used appropriately to account for the overall underlying mood and attitude expressed in the poem, whether directed to sadness or to happiness. This has been done following traditional approaches which assume that the underlying feeling of a poem is strictly related to the sounds conveyed by the words besides/beyond their meaning. In that study we used part of Shakespeare's Sonnets. We have now extended the analysis to the whole collection of 154 sonnets, gathering further evidence of the colour-sound-mood relation. We have also extended the semantic-pragmatic analysis to verify hypotheses put forward by other quantitative computationally-based analysis and compare that with our own. In this case, the aim is trying to discover what features of a poem characterize most popular sonnets.

Keywords Computational Linguistic Analysis, Semantics and Pragmatics, Shakespeare's Sonnets

1. Introduction

The contents of a poem cover many different fields from a sensorial point of view to a mental and a auditory linguistic one. A poem may please our hearing for its rhythm and rhyme structure, or simply for the network of alliterations it evokes by its consonances and assonances; it may attract our attention for its structure of meaning, organized on a coherent lattice of anaphoric and coreferential links or suggested and extracted from inferential and metaphorical links to symbolic meanings obtained by a variety of

rhetorical devices. Most if not all of these facets of a poem are derived from the analysis of SPARSAR, the system for poetry analysis which has been presented to a number of international conferences [1,2,3] - and to Demo sessions in its TTS "expressive reading" version [4,5,6]¹.

Most of a poem's content can be captured considering three basic levels or views on the poem itself: one that covers what can be called the overall sound pattern of the poem - and this is related to the phonetics and the phonology of the words contained in the poem - *Phonetic Relational View*. Another view is the one that captures the main poetic devices related to rhythm, that is the rhyme structure and the metrical structure - this view is called *Poetic Relational View*. Finally, there are semantic and pragmatic contents of the poem which are represented by relations entertained by predicates and arguments expressed in the poem, at lexical semantic level, at metaphorical and anaphoric level - this view is called *Semantic Relational View*. Here we use the three views above and the parameters onto which they are based in order to come up with evidence to prove or disprove two intriguing hypotheses. The first one, the colour-sound-mood relation has been the object of various theories and also poems dedicated to vowels - but see below. J.W.Goethe was the supporter of the union of colour and mood, a theory purported in a book devoted to demonstrate the psychological impact of different colours - "Zur Farbenlehre" / Theory of Colours published in 1810². Unlike physicists like Newton, Goethe's concern was not so much with the analytic treatment of colour, as with how colours are perceived and their impact on the psyche and mood. Colours in our case are produced by the system and used to enforce the feelings induced by the sounds of the words composing the poem: they are not in themselves related to meaning. Different facets of a poem are visualized by graphical output of the system and have been implemented by extracting various properties and features of the poem analysed in eleven separate poetic maps. These maps are organized as follows:

- *A General Description map* including seven Macro Indices with a statistical evaluation of such descriptors

¹ The system is now freely downloadable from sparsar.wordpress.com

² Taken from Wikipedia's page on Goethe, p.1.

- as: Semantic Density Evaluation; General Poetic Devices; General Rhetoric Devices etc., Prosodic Distribution; Rhyming Schemes; Metrical Structure;
- Phonetic Relational Views: five maps,
 - Assonances, i.e. all vowels contained in stressed vowel nuclei which have been repeated in the poem within a certain interval – not just in adjacency;
 - Consonances, i.e. all consonant onsets of stressed syllables again repeated in the poem within a certain interval;
 - All word repetitions, be it stressed or unstressed;
 - one for the Unvoiced/Voiced opposition as documented in syllable onset of stressed words (stress demotion counts as unstressed);
 - another one for a subdivision of all consonant syllable onsets, including consonant cluster onsets, and organized in three main phonological classes:
 - Continuants (only fricatives);
 - Obstruents (Plosives and Affricates);
 - Sonorants (Liquids, Vibrants, Approximants; Glides; Nasals).
 - Poetic Relation Views:
 - Metrical Structure, Rhyming Structure and Expected Acoustic Length, all in one single map.
 - Semantic Relational View: four maps,
 - A map including polarity marked words (Positive vs Negative) and words belonging to Abstract vs Concrete semantic class³;
 - A map including polarity marked words (Positive vs Negative) and words belonging to Eventive vs State semantic class;
 - A map including Main Topic words; Anaphorically linked words; Inferentially linked words; Metaphorically linked words i.e. words linked explicitly by “like” or “as”, words linked by recurring symbolic meanings (woman/serpent or woman/moon or woman/rose);
 - A map showing predicate argument relations intervening between words, marked at core argument words only, indicating predicate and semantic role; eventive anaphora between verbs.

Graphical maps highlight differences using colours. The use of colours associated to sound in poetry has a long tradition. Rimbaud composed a poem devoted to “Vowels” where colours were specifically associated to each of the main five vowels. Roman Jakobson wrote extensively about sound and colour in a number of papers [7; 8:188], lately Mazzeo [9]. As Tsur [10] notes, Fónagy [11] wrote an article in which he connected explicitly the use of certain types of consonant sound associated to certain moods: unvoiced and obstruent consonants are associated with aggressive mood;

sonorants with tender moods. Fónagy mentioned the work of M.Macdermott [12] who in her study identified a specific quality associated to “dark” vowels, i.e. back vowels, that of being linked with dark colours, mystic obscurity, hatred and struggle.

As a result, we are using darker colours to highlight back and front vowels as opposed to low and middle vowels, the latter with light colours. The same applies to representing unvoiced and obstruent consonants as opposed to voiced and sonorants. But as Tsur [10:15] notes, this sound-colour association with mood or attitude has no real significance without a link to semantics. In the Semantic Relational View, we are using dark colours for Concrete referents vs Abstract ones with lighter colours; dark colours also for Negatively marked words as opposed to Positively marked ones with lighter colours. The same strategy applies to other poetic maps: this technique has certainly the good quality of highlighting opposing differences at some level of abstraction⁴.

The usefulness of this visualization is intuitively related to various potential users and for different purposes. First of all, translators of poetry would certainly benefit from the decomposition of the poem and the fine-grained analysis, in view of the need to preserve as much as possible of the original qualities of the source poem in the target language. Other possible users are literary critics and literature teachers at various levels. Graphical output is essentially produced to allow immediate and direct comparison between different poems and different poets – see section below for comparisons.

In order to show the usefulness and power of these visualizations, I have chosen two different English poets in different time periods: Shakespeare with Sonnet 1 and Sylvia Plath, with Edge.

The second hypothesis concerns popularity that some of the Sonnets have achieved in time and the possibility that artistic creativity in poetry is highlighted by extracting appropriate parameters. The study should set apart intrinsic properties of famous sonnets, those features of success that should guarantee a poem to stand out. We will use parameters produced and highlighted in a previous section, and will concentrate on differences characterizing a subset of the whole collection, including the most famous ones, sonnets 18, 29, 30, 73, 116, 126 and 130, but not only.

The chapter is organized as follows: a short state of the art in the following section; then the system SPARSAR is presented with the views of two poems accompanied by comments; data derived from the analysis that will be used to prove or disprove our two hypotheses; some conclusion.

³ see in particular Brysbaert et al. 2014 that has a database of 40K entries. We are also using a manually annotated lexicon of 10K entries and WordNet supersenses. We are not using MRCDatabase which only has some 8,000 concrete + some 9,000 imagery classified entries because it is difficult to adapt and integrate into our system.

⁴ our approach is not comparable to work by Saif Mohammad [13,14], where colours are associated to words on the basis of what their mental image may suggest to the mind of annotators hired via Mechanical Turk. The resource only contains word-colour association for some 12,000 entries over the 27K items listed. It is however comparable to a long list of other attempts at depicting phonetic differences in poems as will be discussed in the next section.

2. Related Work

Computational work on poetry addresses a number of subfields which are however strongly related. They include automated annotation, analysis, or translation of poetry, as well as poetry generation, that we comment here below. Other common subfields regard automatic grapheme-to-phoneme translation for out of vocabulary words as discussed in [15,16] use CMU pronunciation dictionary to derive stress and rhyming information, and incorporate constraints on meter and rhyme into a machine translation system. There has also been some work on computational approaches to characterizing rhymes [17] and global properties of the rhyme network (see [18]) in English. Eventually, graphical visualization of poetic features.

Green et al.[19] use a finite state transducer to infer syllable-stress assignments in lines of poetry under metrical constraints. They contribute variations similar to the schemes below, by allowing an optional inversion of stress in the iambic foot. This variation is however only motivated by heuristics, noting that "poets often use the word 'mother' (S* S) at the beginnings and ends of lines, where it theoretically should not appear." So eventually, there is no control of the internal syntactic or semantic structure of the newly obtained sequence of feet: the optional change is only positionally motivated. They employ statistical methods to analyze, generate, and translate rhythmic poetry. They first apply unsupervised learning to reveal word-stress patterns in a corpus of raw poetry. They then use these word-stress patterns, in addition to rhyme and discourse models, to generate English love poetry. Finally, they translate Italian poetry into English, choosing target realizations that conform to desired rhythmic patterns. They, however, concentrate on only one type of poetic meter, the iambic pentameter. What's more, they use the audio transcripts - made by just one person - to create a syllable-based word-stress gold standard corpus for testing, made of some 70 lines taken from Shakespeare's sonnets. Audio transcripts without supporting acoustic analysis⁵ is not always the best manner to deal with stress assignment in syllable positions which might or might not conform to a strict sequence of iambs. There is no indication of what kind of criteria have been used, and it must be noted that the three acoustic cues may well not be congruent (see [20]). So eventually results obtained are rather difficult to evaluate. As the authors note, spoken recordings may contain lexical stress reversals and archaic pronunciations⁶. Their conclusion is that "this useful information is not available in typical pronunciation dictionaries". Further on, (p. 531) they comment "the

probability of stressing 'at' is 40% in general, but this increases to 91% when the next word is 'the'." We assume that demoting or promoting word stress requires information which is context and syntactically dependent. Proper use of one-syllable words remains tricky. In our opinion, machine learning would need much bigger training data than the ones used by the authors for their experiment.

There's an extended number of papers on poetry generation starting from work documented in a number of publications by P. Gervas [21,22] who makes use of Case Based Reasoning to induce the best line structure. Other interesting attempts are by Toivanen et al.[23] who use a corpus-based approach to generate poetry in Finnish. Their idea is to contribute knowledge needed in content and form by means of two separate corpora, one providing semantic content, and another grammatical and poetic structure. Morphological analysis and synthesis is used together with text-mining methods. Basque poetry generation is the topic of Agirrezabal et al. [24] paper which uses POS-tags to induce the linear ordering and WordNet to select best semantic choice in context.

Manurung et al., [25,26] have explored the problem of poetry generation under some constraints using machine learning techniques. With their work, the authors intended to fill the gap in the generation paradigm, and "to shed some light on what often seems to be the most enigmatic and mysterious forms of artistic expression". The conclusion they reach is that "despite our implementation being at a very early stage, the sample output succeeds in showing how the stochastic hillclimbing search model manages to produce text that satisfies these constraints." However, when we come to the evaluation of metre we discover that they base their approach on wrong premises. The authors quote the first line of what could be a normal limerick but get the metrical structure totally wrong. In limericks, what we are dealing with are not dactyls - TAtata - but anapests, tataTA, that is a sequence of two unstressed plus a closing stressed syllable. This is a well-known characteristic feature of limericks and the typical rhythm is usually preceded and introduced by a iamb "there ONCE", and followed by two anapests, "was a MAN", "from maDRAS". Here in particular it is the syntactic-semantic phrase that determines the choice of foot, and not the scansion provided by the authors⁷.

Reddy & Knight [27] produce an unsupervised machine learning algorithm for finding rhyme schemes which is intended to be language-independent. It works on the intuition that "a collection of rhyming poetry inevitably contains repetition of rhyming pairs. ... This is partly due to sparsity of rhymes – many words that have no rhymes at all, and many others have only a handful, forcing poets to reuse rhyming pairs." The authors harness this repetition to build an unsupervised algorithm to infer rhyme schemes, based on

5 One questions could be "Has the person transcribing stress pattern been using pitch as main acoustic correlate for stress position, or loudness (intensity or energy) or else durational patterns?". The choice of one or the other acoustic correlated might change significantly the final outcome.

6 At p.528 they present a table where they list a number of words - partly function and partly content words - associated to probability values indicating their higher or lower propensity to receive word stress. They comment that "Function words and possessives tend to be unstressed, while content words tend to be stressed, though many words are used both ways".

7 "For instance, the line 'There /once was a /man from Ma/dras', has a stress pattern of (w,s,w,w,s,w,w,s). This can be divided into feet as (w),(s,w,w),(s,w,w),(s). In other words, this line consists of a single upbeat (the weak syllable before the first strong syllable), followed by 2 dactyls (a classical poetry unit consisting of a strong syllable followed by two weak ones), and ended with a strong beat."(ibid.7)

a model of stanza generation. We test the algorithm on rhyming poetry in English and French.” The definition of rhyme the authors used is the strict one of perfect rhyme: two words rhyme if their final stressed vowels and all following phonemes are identical. So no half rhymes are considered. Rhyming lines are checked from CELEX phonological database[28].

There's a small number of rule-based systems available for download which need to be considered before presenting our system, and they are – listed from the oldest to the latest:

- the *Scandroid* by C.Hartman (2004/5), downloadable at <http://oak.conncoll.edu/cohar/Programs.htm>, and presented in [29]
- the *Stanford Literary Lab* by Algee-Hewitt, M., Heuser, R. Kraxenberger, M., Porter, J., Sensenbaugh, J., and Tackett, J. (2014), downloadable at <https://github.com/quadrismegistus/litlab-poetry>, and presented in [30,31]
- the University of Toronto Canada *Representative Poetry Online* project carried out by M.R. Plamondon and documented in [32], downloadable at Library website <http://rpo.library.utoronto.ca/>
- a collaborative effort carried out by American and German universities called *MYOPIA*, presented in [33] and available at two websites by the main author Helen Armstrong, <https://lecture2go.uni-hamburg.de/konferenzen/-/k/13930>, <http://www.helenarmstrong.us/design/myopia/>
- *ZeusScansion* for the scansion of English poetry by M. Agirrezabal et al. presented in [34], and available at <https://github.com/manexagirrezabal/zeuscansion>
- *RhymeDesign* a tool designed for the analysis of metric and rhythmic devices, by N.McCurdy et al. [35], a tool previously called *Poemage*, documented at <http://www.sci.utah.edu/~nmccurdy/Poemage/> and now presented as project at http://ninamccurdy.com/?page_id=398

A number of more or less recent works have addressed the problem related to rhyme identification, by Manish Chaturvedi et al. [36] and by Karteek Addanki and Dekai Wu [37], but also previously by Hussein Hirjee and Daniel Brown [38] and by Susan Bartlett et al. [39]. Eventually a selected list of authors have specifically addressed the problem of visualization of linguistic and literary data in recent and not so recent works, notably on poetry visualization [40], literary analysis and concordancing [41,42,43].

3. SPARSAR - Automatic Analysis of Poetic Structure and Rhythm

SPARSAR produces a deep analysis of each poem at different levels: it works at sentence level at first, then at verse level and finally at stanza level (see Figure 1 below). The structure of the system is organized as follows: at first syntactic, semantic and grammatical functions are evaluated.

Then the poem is translated into a phonetic form preserving its visual structure and its subdivision into verses and stanzas. Phonetically translated words are associated to mean duration values taking into account stressed syllables and syllable position in the word. At the end of analysis the system can measure the following parameters: mean verse length in terms of msec. and in number of feet. The latter is derived by a verse representation of metrical structure after scansion. The other important component of the analysis of rhythm is constituted by the algorithm that measures and evaluates rhyme schemes at stanza level, and then computes the overall rhyming structure at poem level. As regards syntax, we now have at our disposal, chunks and dependency structures if needed, which have been computed by the tagger and parser of the previous modules. To complete our analysis, we introduce semantics both in the version of a classifier and by isolating each verbal complex. In this way we verify propositional properties like presence of negation; we compute subjectivity and factuality from a crosscheck with modality, aspectuality – that we derive from our lexica – and tense (more on this topic below). On the other hand, the classifier has two different tasks: distinguishing concrete from abstract nouns, identifying highly ambiguous from singleton concepts: this is computed from number of possible meanings defined in WordNet and other similar repositories. Eventually, we carry out a sentiment analysis of every poem, thus contributing a three-way classification: neutral, negative, positive that can be used as a powerful tool for evaluation purposes. Kao & Jurafsky[44] who also used the tool denounces that. In that paper, Jurafsky works on the introduction of a semantic classifier to distinguish concrete from abstract nouns. More about semantics in a section below.

In building our system, we have been inspired by Kaplan's tool APSA[45,46], and started developing a similar system, but which was more transparent and more deeply linguistically-based. The main new target in our opinion, had to be an index strongly semantically based, i.e. a “Semantic Density Index” (SDI). With this definition we now refer to the idea of classifying poems according to their intrinsic semantic density in order to set apart those poems which are easy to understand from those that require a rereading and still remain somewhat obscure. An intuitive notion of SDI can be formulated as follow:

- easy to understand are those semantic structures which contain a proposition, made of a main predicate and its arguments
- difficult to understand are on the contrary semantic structures which are filled with nominal expressions, used to reinforce a concept and are juxtaposed in a sequence
- also difficult to understand are sequences of adjectives and nominals used as modifiers, union of such items with a dash.

Other elements that we introduce in the definition of semantic parameters are presence of negation and modality:

this is why we compute Polarity and Factuality. Additional features are obtained by measuring the level of affectivity by means of sentiment analysis, focussing on presence of negative items which contribute to make understanding more difficult.

The Semantic Density Index is derived from the computation of a number of features, some of which have negative import while others positive import. At the end of the computation the index may end up to be positive if the poem is semantically “light”, that is easy to read and understand; otherwise, it is computed as “heavy” which implies that it is semantically difficult.

At the end we come up with a number of evaluation indices that include: a Constituent Density Index, a Sentiment Analysis Marker, a Subjectivity and Factuality Marker. We also compute a Deep Conceptual Index, see below. All these indices contribute to creating the SDI mentioned above.

The procedure is based on the tokenized sentence, which is automatically extracted and may contain many verses up to a punctuation mark, usually period. Then I use the functional structures which are made of a head and a constituent which are measured for length in number of tokens. A first value of SDI comes from the proportion of verbal compounds and non-verbal ones. I assume that a “normal” distribution for a sentence corresponds to a semantic proposition that contains one verbal complex with a

maximum of four non verbal structures. More verbal compounds contribute to reducing the SDI.

The other contribution comes from lemmatization and the association of a list of semantic categories, general semantic classes coming from WordNet or other similar computational lexica. These classes are also called supersense classes. As a criterion for grading difficulty, I consider more difficult to understand a word which is specialized for a specific semantic domain and has only one such supersense label. On the contrary, words or concepts easy to understand are those that are ambiguous between many senses and have more semantic labels associated to the lemma. A feature derived from quantitative linguistic studies is the rare words, which are those words that appear with less than 4 occurrences in frequency lists. I use the one derived from Google GigaWord.

The index has a higher value for those cases of high density and a lower value for the contrary. It is a linear computation and includes the following features: the ratio of number of words vs number of verbs; the ratio of number of verbal compounds vs non-verbal ones; the internal composition of non-verbal chunks: every additional content word increases their weight (functional words are not counted); the number of semantic classes. Eventually a single index is associated to the poem which should be able to differentiate those poems which are easy from the cumbersome ones.

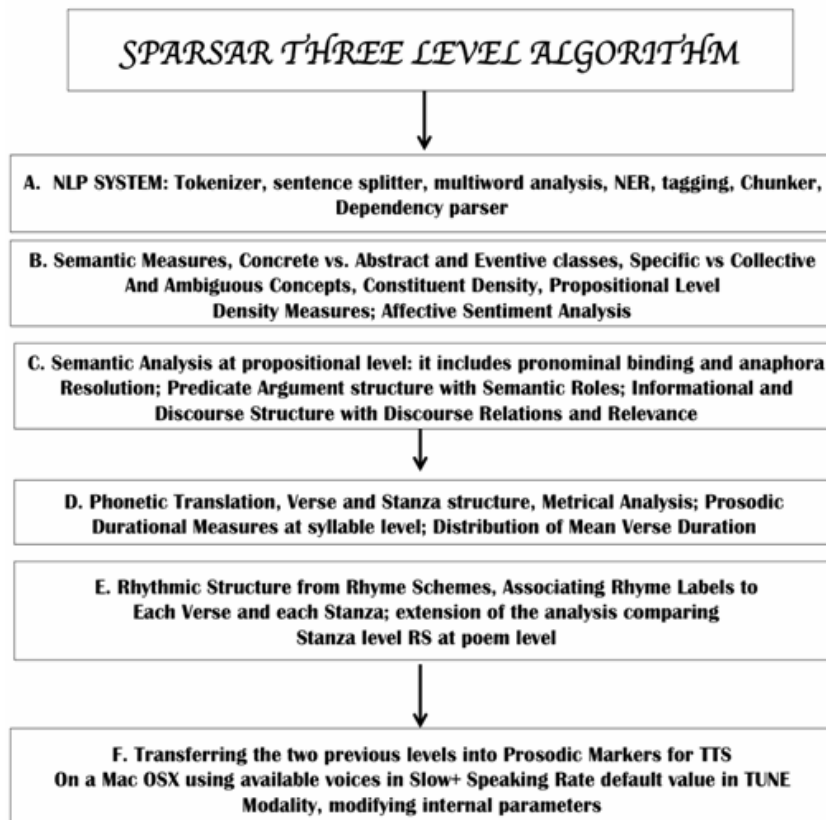


Figure 1. The SPARSAR three-level system

What we do is dividing each item by the total number of tagged words and of chunks. In detail, we divide verbs found by the total number of tokens (the more the best); we divide adjectives found by the total number of tokens (the more the worst); we divide verb structures by the total number of chunks (the more the best); we divide inflected vs uninflected verbal compounds (the more the best); we divide nominal chunks rich in components : those that have more than 3 members (the more the worst); we divide semantically rich (with less semantic categories) words by the total number of lemmas (the more the worst); we count rare words (the more the worst); we count generic or collective referred concepts (the more the best); we divide specific vs ambiguous semantic concepts (those classified with more than two senses) (the more the worst); we count doubt and modal verbs, and propositional level negation (the more the worst); we divide abstract and eventive words vs concrete concepts (the more the worst); we compute sentiment analysis with a count of negative polarity items (the more the worst).

Another important index we implemented is the Deep Conceptual index, which is obtained by considering the proportion of Abstract vs Concrete words contained in the poem. This index is then multiplied with the Propositional Semantic Density which is obtained at sentence level by computing how many non verbal, and amongst the verbal, how many non inflected verbal chunks there are in a sentence – more on these indices below.

3.1. Computing Metrical Structure and Rhyming Scheme

Any poem can be characterized by its rhythm which is also revealing of the poet's peculiar style. In turn, the poem's rhythm is based mainly on two elements: meter, that is distribution of stressed and unstressed syllables in the verse, presence of rhyming and other poetic devices like alliteration, assonance, consonance, enjambments, etc. which contribute to poetic form at stanza level. This level is combined then with syntax and semantics to produce the adequate breath-groups and consequent subdivision: these usually coincide with line-end words, but they may continue to the following line through enjambments.

A poetic foot can be marked by a numerical sequence as for instance in Hayward[47,48] where 0/1 is used: “0” for unstressed and “1” for stressed syllables to feed a connectionist model of poetic meter from a manually transcribed corpus. We also use sequences of 0/1 to characterize poetic rhythm. But then we deepen our analysis by considering stanzas as structural units in which rhyming plays an essential role. What is paramount in our description of rhythm, is the use of the acoustic parameter of duration. The use of acoustic duration allows our system to produce a model of a poetry reader that we implement by speech synthesis. The use of objective prosodic rhythmic and stylistic features, allows us to compare similar poems of the same poet and of different poets both prosodically and metrically.

To this aim we assume that syllable acoustic identity changes as a function of three parameters:

- internal structure in terms of onset and rhyme which is characterized by number of consonants, consonant clusters, vowel or diphthong
- position in the word, whether beginning, end or middle
- primary stress, secondary stress or unstressed

These data have been collected in a database called VESD (see [49,50]).

The analysis starts by translating every poem into its phonetic form - see Figure 2. below. After reading out the whole poem on a line by line basis and having produced all phonemic transcription, we look for poetic devices. Here assonances, consonances, alliterations and rhymes are analysed and then evaluated. Then we compute metrical structure, which is the alternation of beats: this is computed by considering all function or grammatical words which are monosyllabic as unstressed – in fact not all of them, heavy monosyllabic function words are left stressed, i.e. “through”. We associate a “0” to all unstressed syllables, and a value of “1” to all stressed syllables, thus including both primary and secondary stressed syllables – only if needed. We try to build syllables starting from longest possible phone sequences to shortest one. This is done heuristically trying to match pseudo syllables with our syllable list. Matching may fail and will then result in a new syllable which has not been previously met. We assume that any syllable inventory is deficient, and is never sufficient to cover the whole spectrum of syllables available in the English language.

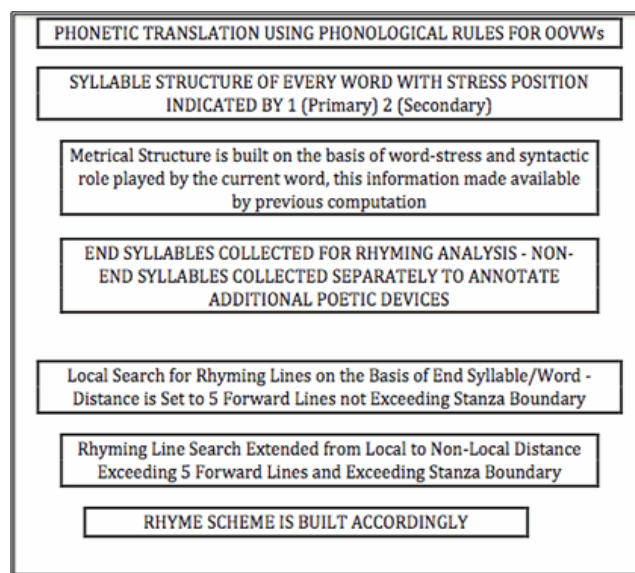


Figure 2. SPARSAR Poetic Analyzer

For this reason, we introduced a number of phonological rules to account for any new syllable that may appear. To produce our prosodic model we take mean durational values. We also select, whenever possible, positional and stress values. We also take advantage of syntactic information computed separately to highlight chunks' heads as produced by our bottom-up parser. In that case, stressed syllables take

maximum duration values. Dependent words on the contrary are “demoted” and take minimum duration values.

On a second pass we check for sequences of zeros and ones in order to demote/promote syllables that require it, after having counted total number of syllables per line. This is done to adjust metrical structure. We do this recursively by searching sequences made of three zeroes – 0,0,0 – or three ones – 1,1,1. In both cases we have at our disposal tagging and head information associated to current word. If none of the two are available we modify the second or middle value thus transforming the previous sequences into “0,1,0” and “1,0,1”, respectively.

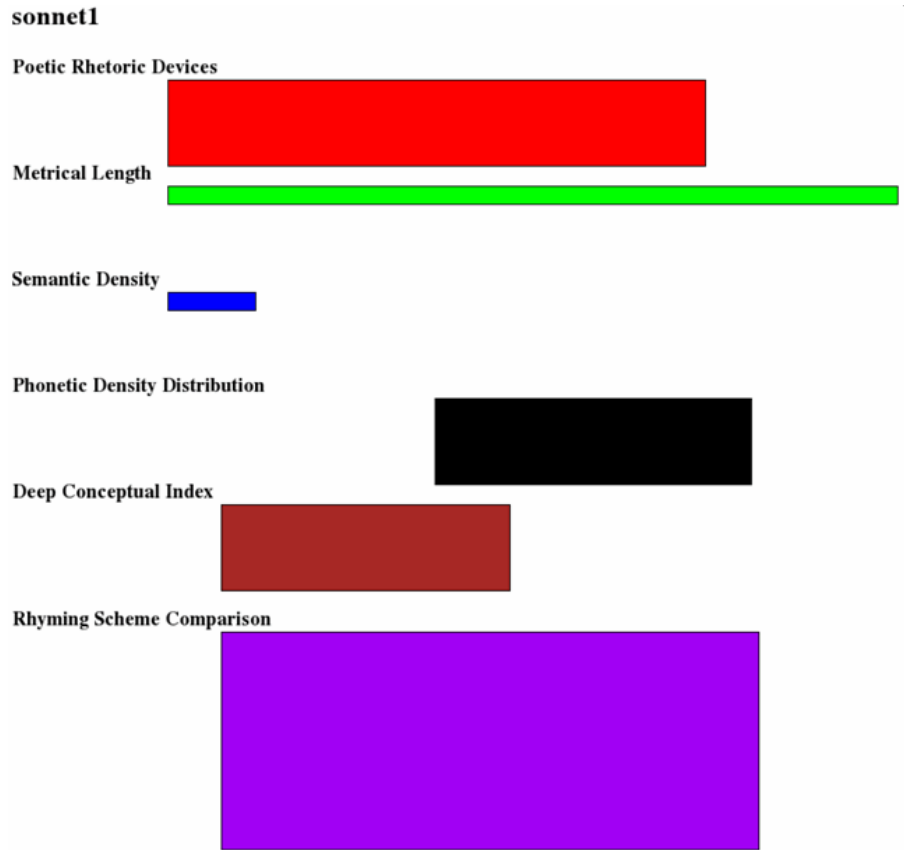
Durations are then collected at stanza level and a statistics is produced. Metrical structure is used to evaluate statistical measures of its distribution in the poem. As can be easily gathered, it is difficult to find lines with identical number of syllables, identical number of metrical feet and identical metrical verse structure. If we consider the sequence “01” as representing the typical iambic foot, and the iambic pentameter as the typical verse metre of English poetry, in our transcription it is easy to see that there is no poem strictly respecting it. On the contrary we find trochees, “10”, dactyls, “100”, anapests, “001” and spondees, “11”. At the end of the computation, the system is able to measure two important indices: “mean verse length” and “mean verse length in no. of feet” that is mean metrical structure. Additional measures that we are now able to produce are related to rhyming devices. Since we intended to take into account structural internal rhyming scheme and their persistence in the poem, we enriched our algorithm with additional data. These measures are then accompanied by information derived from two additional component: word repetition and rhyme repetition at stanza level. Sometimes also refrain may apply, that is the repetition of an entire line of verse. Rhyming schemes together with metrical length, are the strongest parameters to consider when assessing similarity between two poems.

Eventually we need to reconstruct the internal structure of metrical devices used by the poet: in some cases, also stanza repetition at poem level may apply. We then use this information as a multiplier. The final score is then tripled in case of structural persistence of more than one rhyming

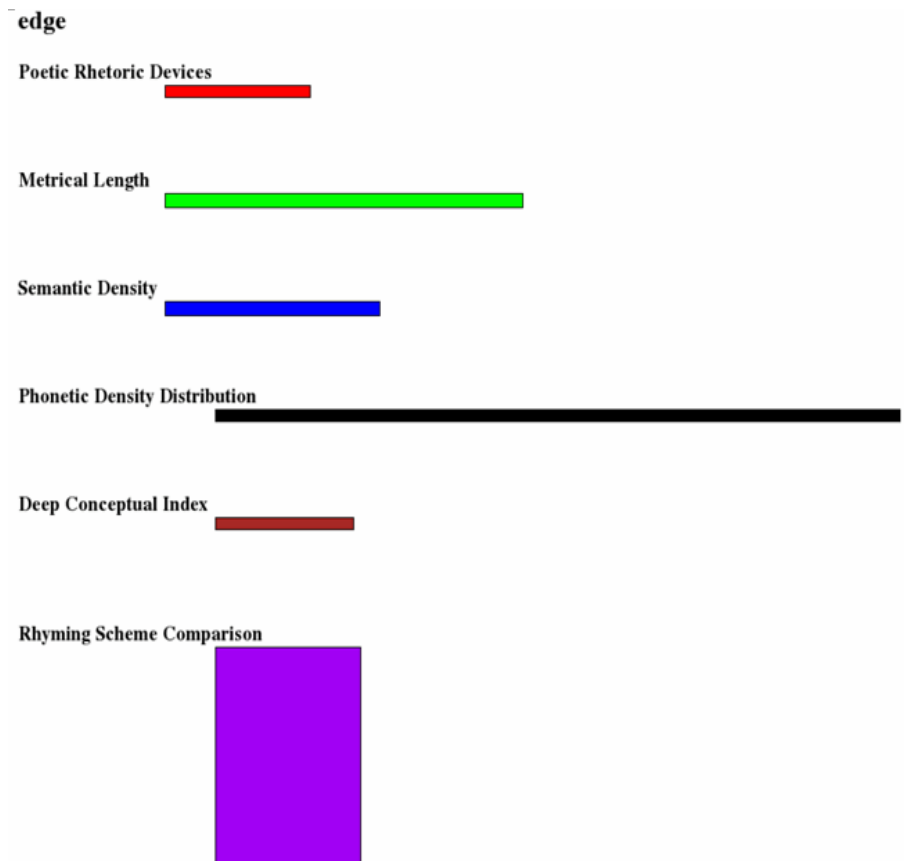
scheme; for only one repeated rhyme scheme, it is doubled. With no rhyming scheme, there is no increase in the linear count of rhetorical and rhyming devices. To create the rhyming scheme we assign labels to each couple of rhyming line and then match recursively each final phonetic word with the following ones, starting from the closest to the one that is further apart. Each time we register the rhyming words and their distance. In the following pass we reconstruct the actual final line numbers and then produce an indexed list of couples, Line Number-Rhyming Line for all the lines, stanza boundaries included. Eventually, we associate alphabetic labels to the each rhyming verse starting from A to Z. A simple alphabetic incremental mechanism updates the rhyme label. This may go beyond the limits of the alphabet itself and in that case, double letters are used.

What is important for final evaluation, is persistence of a given rhyme scheme, how many stanza contain the same rhyme scheme and the length of the scheme. A poem with no rhyme scheme is much poorer than a poem that has at least one, so this needs to be evaluated positively and this is what we do. Rhetorical and rhyming devices are then used, besides semantic and conceptual indices, to match and compare poems and poets.

SPARSAR visualizes differences by increasing the length and the width of each coloured bar associated to the indices. Parameters evaluated and shown by coloured bars include: Poetic Rhetoric Devices (in red); Metrical Length (in green); Semantic Density (in blue); Prosodic Structure Dispersion (in black); Deep Conceptual Index (in brown); Rhyming Scheme Comparison (in purple). Their extension indicates the dimension and size of the index: longer bars are for higher values. In this way it is easily shown which component of the poem has major weight in the evaluation. We show here below the graphical output of this type of evaluation for Sonnet 1 and the poem *Edge* by Sylvia Plath. Differences can be easily appreciated at all levels of computation: coloured bars are longer and wider in Sonnet 1 for Poetic Devices and Metrical Length; they are longer in *Edge* for Semantic Density and Phonetic Density Distribution. Deep Conceptual Index and Rhyming Scheme are both higher for Sonnet 1.



Global Diagram 1. General Description Map for Sonnet1



Global Diagram 2. General Description Map for Edge

Parameters related to the Rhyming Scheme (RS) contribute a multiplier – as said above - to the already measured metrical structure which includes: a count of metrical feet and its distribution in the poem; a count of rhyming devices and their distribution in the poem; a count of prosodic evaluation based on durational values and their distribution. RS is based on the regularity in the repetition of a rhyming scheme across the stanzas or simply the sequence of lines in case the poem is not divided up into stanzas. We don't assess different RSs even though we could: the only additional value is given by the presence of a Chain Rhyme scheme, which is a rhyme present in one stanza which is inherited by the following stanza. Values to be computed are related to the Repetition Rate (RR), that is how many rhymes are repeated in the scheme or in the stanza: this is a ratio between number of verses and their rhyming types. For instance, a scheme like AABCC, has a higher repetition rate (corresponding to 2) than say AABCDD (1.5), or ABCDD (1.5). The RR is a parameter linked to the length of the scheme, but also to the number of repeated schemes in the poem: RS may change during the poem and there may be more than one scheme. A higher evaluation is given to full rhymes, which add up the number of identical phones, with respect to half-rhymes and other types of rhyme which on the contrary count only half that number. We normalize final

evaluation to balance the difference between longer vs. shorter poems, where longer poems are rewarded for the intrinsic difficulty of maintaining identical rhyming schemes with different stanzas and different vocabulary.

4. Three Views via Poetic Graphical Maps

The basic idea underlying poetic graphical maps is that of making available to the user an insight of the poem which is hardly realized even if the analysis is carried out manually by an expert literary critic. This is also due to the fact that the expertise required for the production of all the maps ranges from acoustic phonetics to semantics and pragmatics, a knowledge that is not usually possessed by a single person. All graphical representations associated to the poems are produced by Prolog SWI, inside the system which is freely downloadable from its website, at sparsar.wordpress.com. For lack of space, we show maps related to one of Shakespeare's Sonnets, Sonnet 1 and compare it to Sylvia Plath's Edge, to highlight similarities and to show that the system can handle totally different poems still allowing comparisons to be made neatly.

Poem and Poetic Devices :- Assonance Alliterations sonnet_1

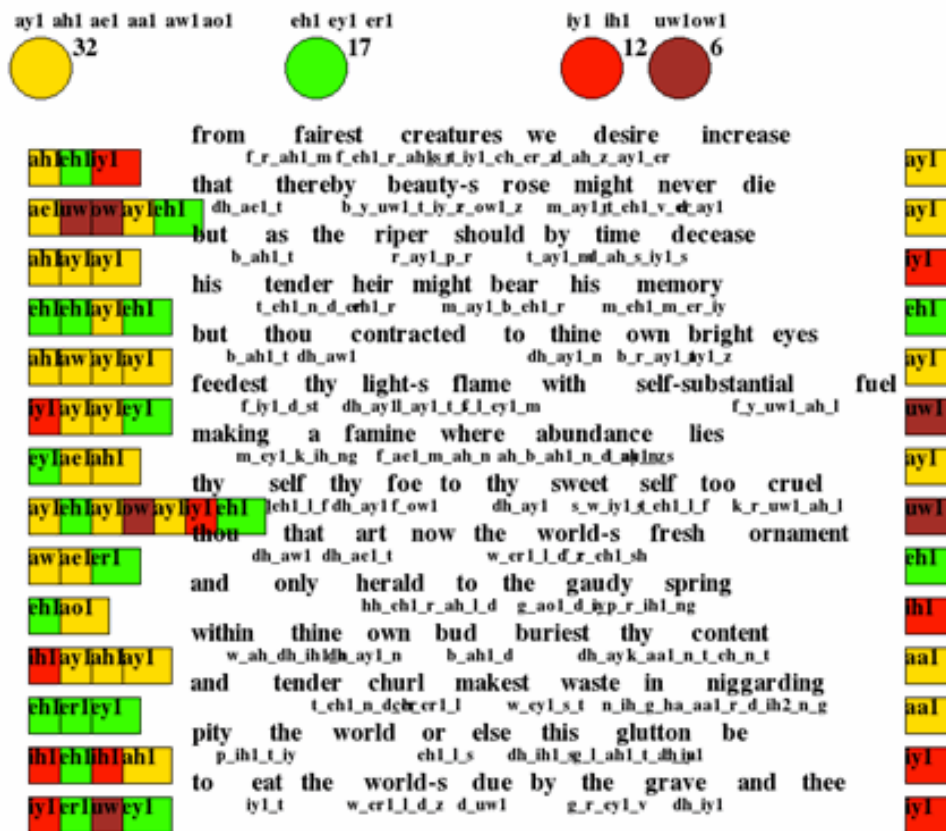


Diagram 1. Assonances

Poem and Poetic Devices :- Consonance Alliterations
sonnet1



Diagram 2. Consonances

Poem and Poetic Devices :- Phonetic Voiced/Unvoiced Map
sonnet_1

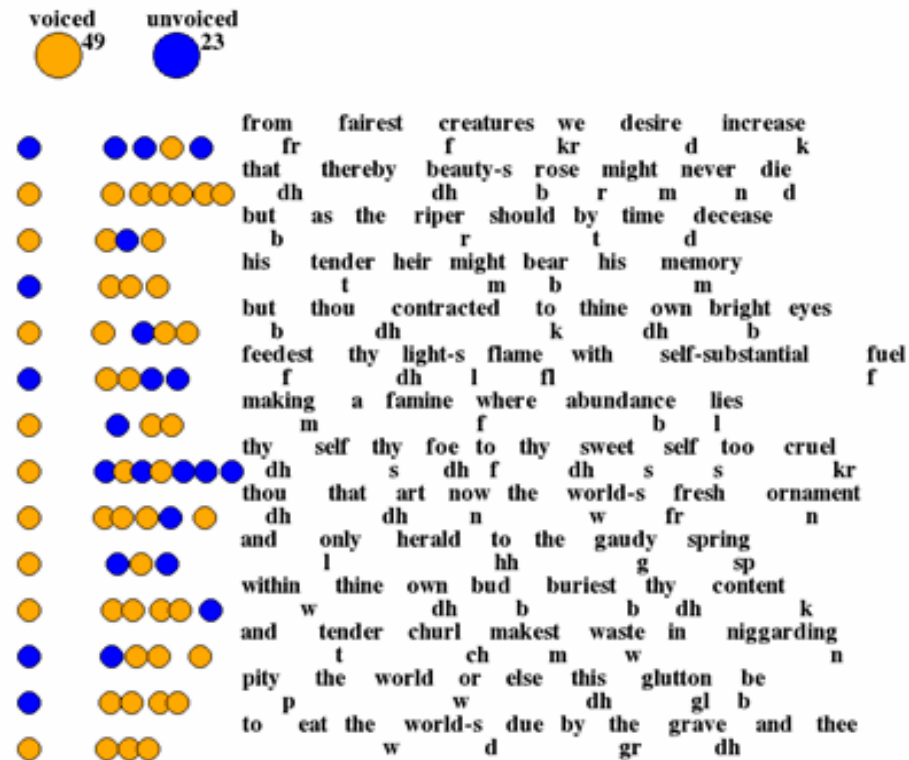


Diagram 3. Voiced/Unvoiced

Poem and Rhythm :- Rhyme Structure, Metrical Feet and Acoustic Length
sonnet_1

A	11	from fairest creatures we desire increase f_r_ah_m f_ch r_ah_s_t k_r_ly ch_er_x w_ly d_ah r_ay er ih2_n k_r_ly_s 1 1 0 1 0 1 0 1 0 0 1
F	10	that thereby beauty-s rose might never die dh_ae_t dh_ch_r b_ay b_y_ow_t iy_x r_ow_x m_ay_t n_ch v_er d_ay 1 0 1 1 0 1 1 1 0 1
A	9	but as the riper should by time decrease b_ah_t ae_x dh_ah r_ay_p_r sh_ah_d b_ay t_ay_m d_ah s_ly_s 1 0 0 1 1 0 1 0 1
G	10	his tender heir might bear his memory hh_h_x t_ch_n d_er ch_r m_ay_t b_ch_r hh_h_x m_ch m_er iy 0 1 0 1 1 1 0 1 0 0
B	10	but thou contracted to thine own bright eyes b_ah_t dh_ow k_aa2_n t_r_oe_k t_ah_d t_ow dh_ay_n ow_n b_r_ay_t ay_x 0 1 0 1 0 0 1 1 0 1
C	11	feedest thy light-s flame with self-substantial fuel f_ly_d_st dh_ay t_ay_t_s f_ly_m w_hh dh_s_ch_l_f s_ah_b_s t_ae_n ch_ah_l f_r_ow ah_l 1 0 1 1 0 1 0 1 0 1 0
B	10	making a famine where abundance lies m_ey k_h_ng ah f_ae m_ah_n w_ch_r ah b_ah_n d_ah_n_s t_ay_x 1 0 0 1 0 1 1 0 0 1
C	11	thy self thy foe to thy sweet self too cruel dh_ay s_ch_l_f dh_ay t_ow t_ow dh_ay s_w_ly_t s_ch_l_f t_ow k_r_ow ah_l 0 1 0 1 0 0 1 1 1 1 0
D	10	thou that art now the world-s fresh ornament dh_ow dh_ae_t aa_r_t n_ow dh_ah w_er_l_d_x f_r_ch_sh ae_r n_ah m_ah_n_t 1 1 1 1 0 1 1 1 0 0
H	10	and only herald to the gaudy spring ae_n_d ow_n t_ly hh_ch r_ah_l_d t_ow dh_ah g_ao d_ly s_p_r_h_ng 0 1 0 1 0 0 0 1 0 1
D	10	within thine own bud buriest thy content w_ah dh_h_n dh_ay_n ow_n b_ah_d b_ah_r iy_s dh_ay k_aa2_n t_ch_n_t 0 1 1 0 1 1 0 0 1 0

Diagram 4. Poetic Relational View

We start commenting Phonetic Relational View and its related maps. First map is concerned with Assonances. Here sounds are grouped into Vowel Areas, as said above, which include also diphthongs: now, in area choice what we have considered is the onset vowel. We have disregarded the offset glide which is less persistent and might also not reach its target articulation. We also combine together front high vowels, which can express suffering and pain, with back dark vowels.

Assonances and Consonances are derived from syllable structure in stressed position of repeated sounds within a certain line span: in particular, Consonances are derived from syllable onset while Assonances from syllable nuclei in stressed position. Voiced/Unvoiced from all consonant onsets of stressed words.

As can be noticed from the maps below, the choice of warm colours is selected for respectively, CONTINUANT (yellow), VOICED (orange), SONORANT (green), Centre/Low Vowel Area (gold), Middle Vowel Area (green); and cold colours respectively for UNVOICED (blue), Back Vowel Area (brown). We used then red for OBSTRUENT (red), Front High Vowel Area (red), to indicate suffering and surprise associated to speech signal interruption in obstruents.

The second set of views is the Poetic Relations View. It is obtained by a single graphical map which however condenses five different levels of analysis. The Rhyming

Structure is obtained by matching line endings in their phonetic form. The result is an uppercase letter associated to each line, on the left. This is accompanied by a metrical measure indicating the number of syllables contained in the line. Then the text of the poem appears and underneath each word the phonetic translation at syllable level. Finally, another annotation is added, by mapping syllable type with sequences of 0/1. The additional important layer of analysis that this view makes available is an acoustic phonetic image of each line represented by a coloured streak computed on the basis of the average syllable length in msec derived from our database of syllables of British English – for a similar approach see Tsur [51].

Eventually the third set of views, the Semantic Relational View, produced by the modules of the system derived from VENSES [52]. This view is organized around four separate graphical poetic maps: a map which highlights Event and State words/lemmata in the poem; a map which highlights Concrete vs Abstract words/lemmata. Both these maps address nouns and adjectives. They also indicate Affective and Sentiment analysis (see [53,54]), an evaluation related to nouns and adjective – which however will be given a separate view when the Appraisal-based dictionary will be completed. A map which contains main Topics, Anaphoric and Metaphoric relations, and a final map with Predicate-arguments relations, where Subject and Object arguments are reported again with different colours.

Poem and Semantics :- PredicateArgument Relations and Event Anaphora
sonnet1

SUBJECT-ARG OBJECT-ARG EVENT-ANAPHORA



Diagram 5. Predicate-Argument Relations

Poem and Rhetoric Devices :- Polarity and Abstract/Concrete Words
sonnet1

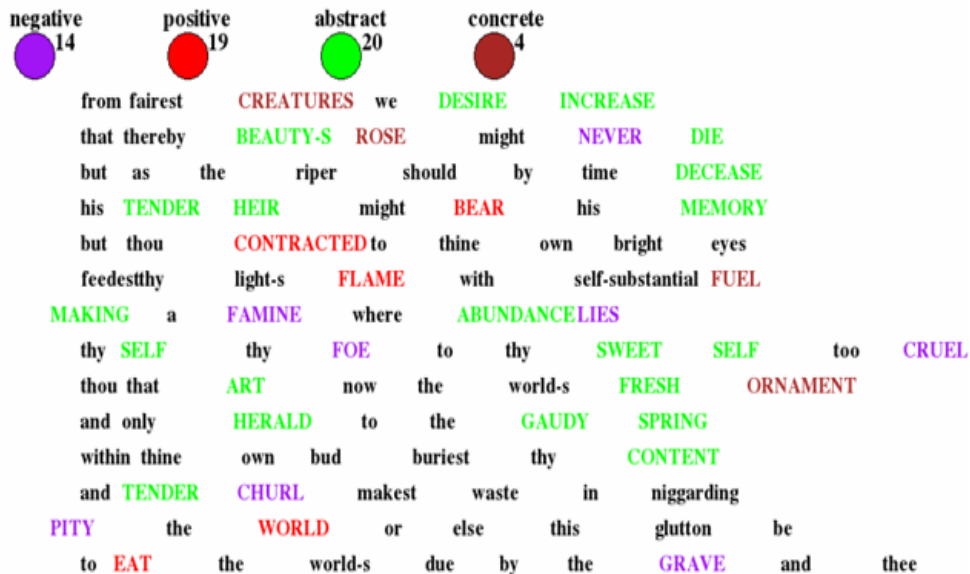


Diagram 6. Abstract/Concrete – Polarity

**Poem and Semantics :- Main Topics and Anaphora
sonnet1**

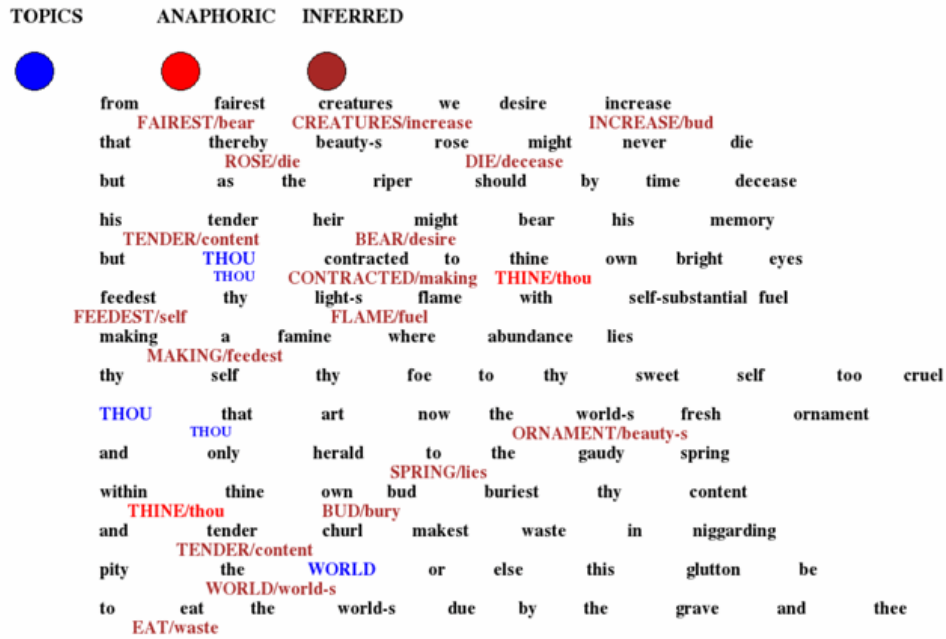


Diagram 7. Main Topics Anaphora and Metaphoric Relations

**Poem and Rhetoric Devices :- Polarity and Events/States Words
sonnet1**



Diagram 8. Events and State - Polarity

In the Phonetic Relations Views for Sonnet 1 reported in Diagram4, the choice of words is strongly related to the main theme and the result is a brighter, clearer more joyful overall sound quality of the poem: number of voiced is the double of unvoiced consonants; in particular, number of obstruents is the same as that of continuants and it is half the sum of sonorants and continuants. As to Assonances, we see that A and E sounds - that is open and middle vowels - constitute the majority of sounds, there is a small presence of back and high front vowels: 18/49, i.e. dark are only one third of light sounds. Eventually, the information coming from affective analysis confirms our previous findings: we see a majority of positive words/propositions, 18/15.

This interpretation of the data is expected also for other poets and is proven by Sylvia Plath’s Edge, a poem the author wrote some week before her suicidal death. It’s a terrible and beautiful poem at the same time: images of death are evoked and explicitly mentioned in the poem, together with images of resurrection and nativity. The poem starts with an oxymoron: “perfected” is joined with “dead body” and both are predicated of the “woman”. We won’t be able to show all the maps for lack of space, but the overall sound pattern is strongly reminiscent of a death toll. In the Consonances map, there’s a clear majority of obstruent sounds and the balance between voiced/unvoiced consonants is in favour of the latter. In the Assonances map we see that dark vowel sounds are more than light or clear sounds 33/30.

Poem and Poetic Devices :- Consonance Alliterations edge



Diagram 9. Consonances

Poem and Poetic Devices :- Phonetic Voiced/Unvoiced Map

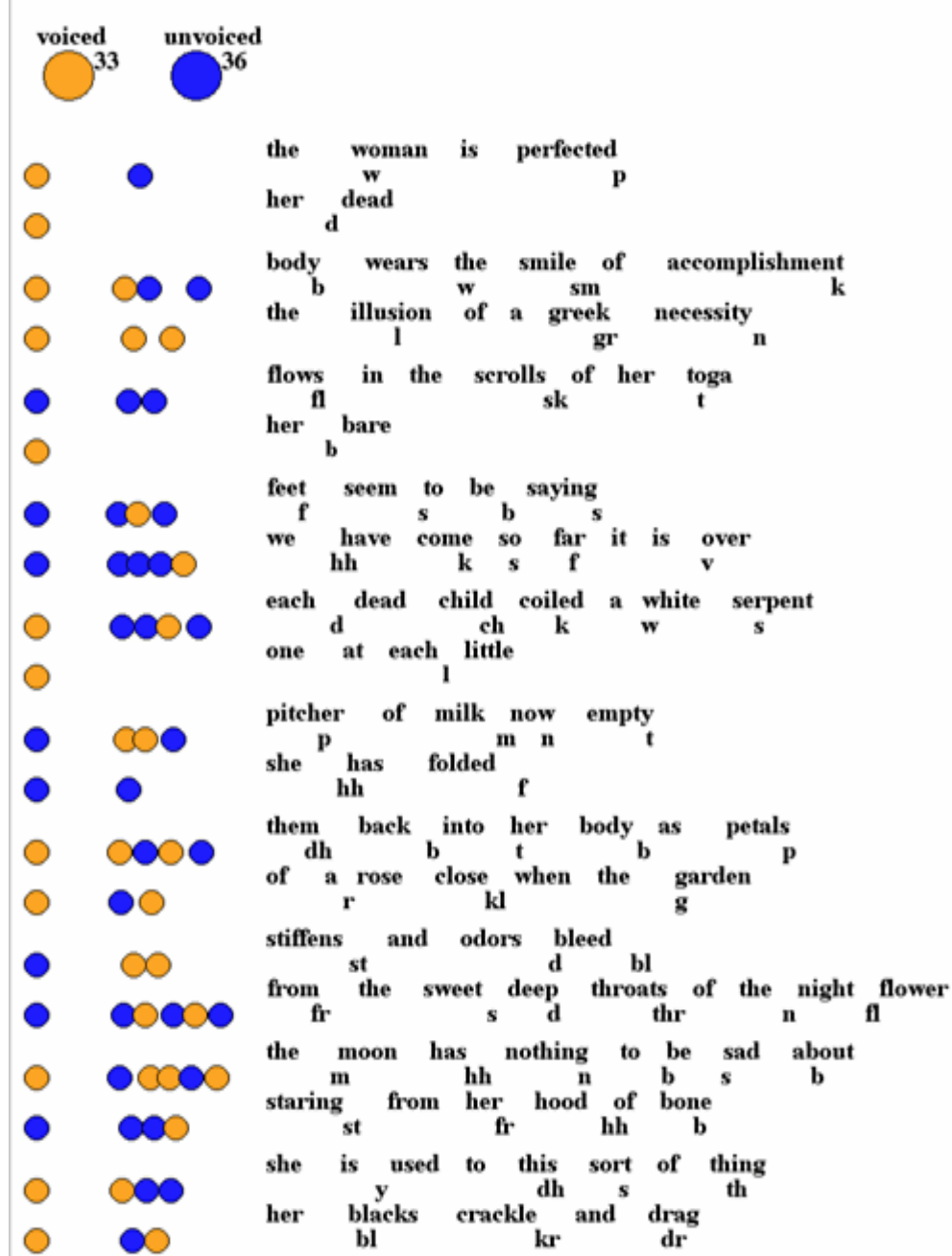


Diagram 10. Voiced/Unvoiced

Poem and Poetic Devices :- Assonance Alliterations edge

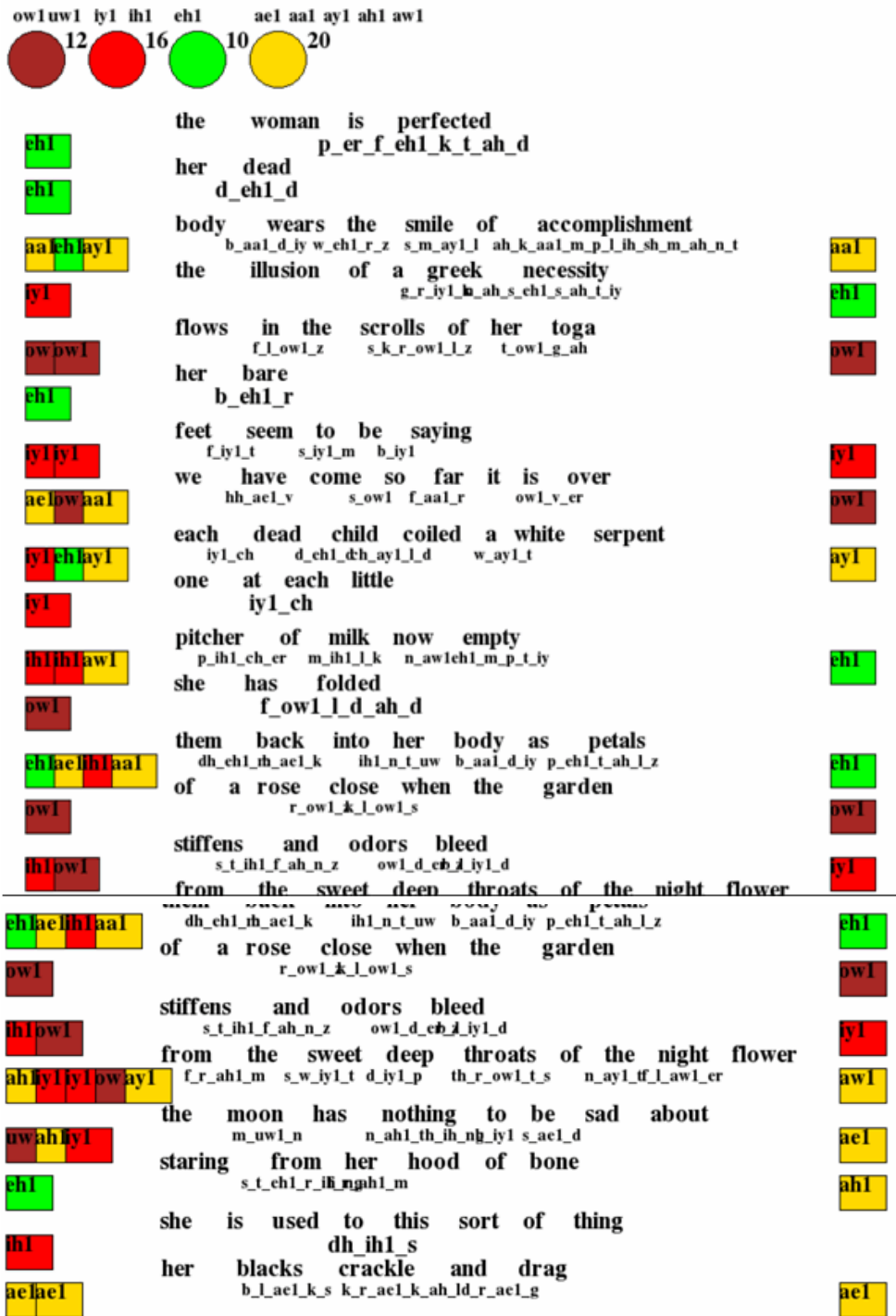


Diagram 11. Assonances Edge

If we look at the topics and the coherence through anaphora, we find that the main topic is constituted by concepts WOMAN, BODY and CHILD. There's also a wealth of anaphoric relations expressed by personal and possessive pronouns which depend on WOMAN. In addition, the system has found metaphoric referential links with such images as MOON GARDEN and SERPENT. In particular the Moon is represented as human - "has nothing to be sad about".

Poem and Semantics :- Main Topics and Anaphora edge

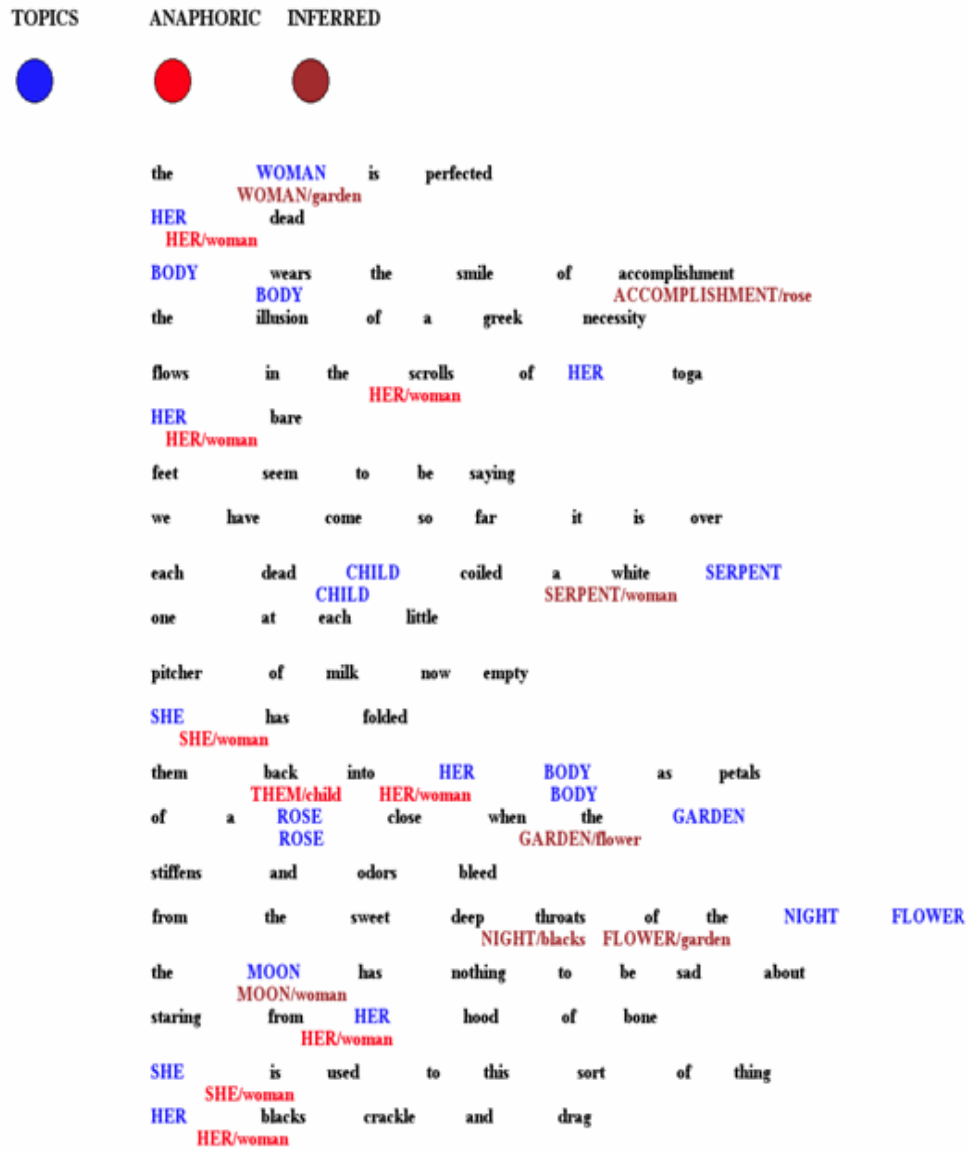


Diagram 12. Topics, Anaphora and Metaphoric Relations

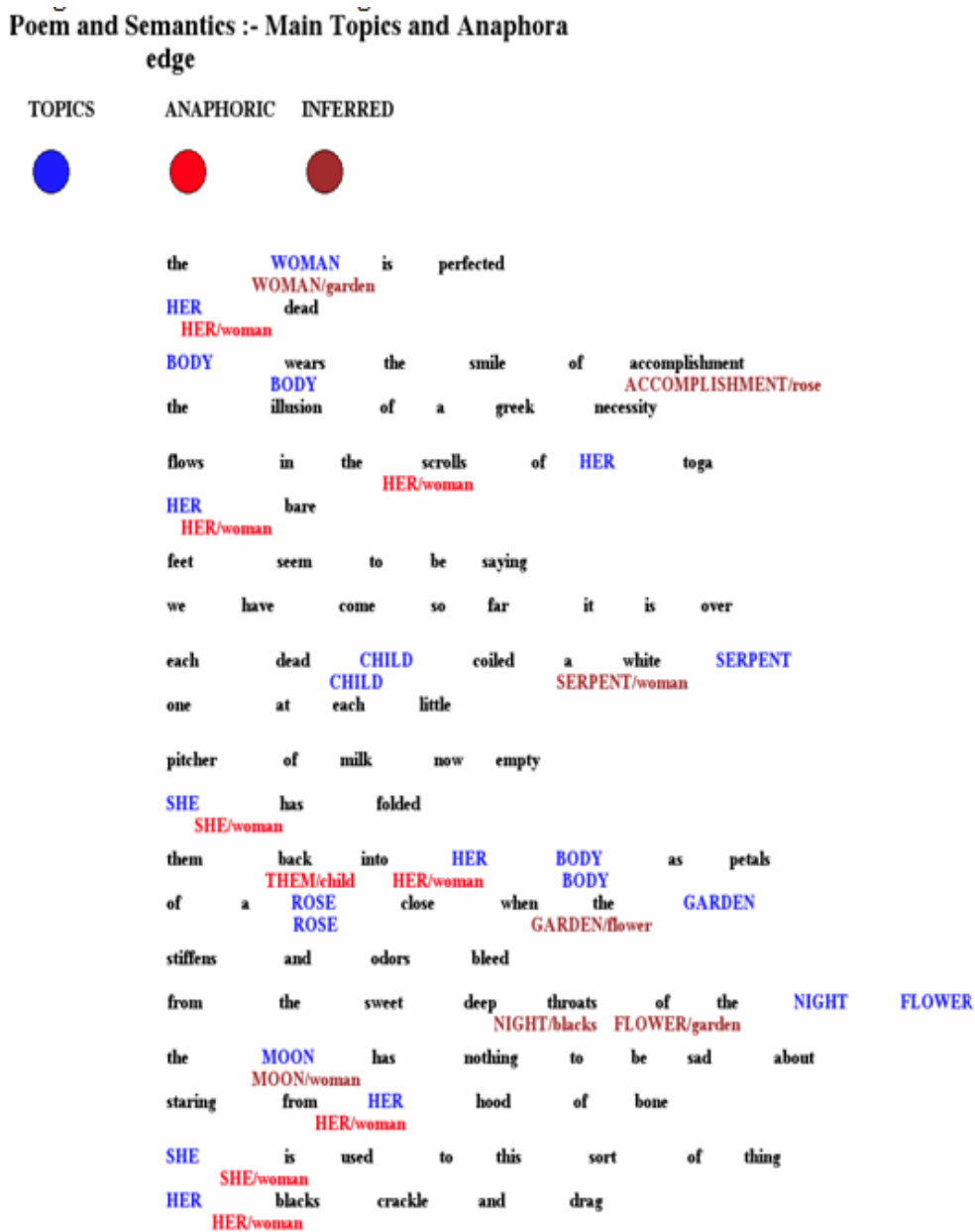


Diagram 13. Abstract/Concrete + Polarity

These images are all possible embodiment of the WOMAN, either directly - the Moon is feminine (she) - or indirectly, when the CHILD that the woman FOLDS the children in her BODY, and the children are in turn assimilated to WHITE SERPENTS. Finally in the Abstract/Concrete map where Polarity is also present we see that Negative items are the majority. Concrete items are also in great amount if compared to other poems.

4.1. Semantic Representations

In this section we will better clarify semantic representations produced by SPARSAR. Semantics in our case refers to predicate-argument structure, negation scope, quantified structures, anaphora resolution and also essentially to propositional level analysis. Propositional

level semantic representation is the basis for discourse structure and discourse semantics contained in discourse relations. It also paves the way for a deep sentiment or affective analysis of every utterance, which alone can take into account the various contributions that may come from syntactic structures like NPs and APs where affectively marked words may be contained. Their contribution needs to be computed in a strictly compositional manner with respect to the meaning associated to the main verb, where negation may be lexically expressed or simply lexically incorporated in the verb meaning itself.

In Fig. 3 we show the architecture of our deep system for semantic and pragmatic processing, in which phonetics, prosodics and NLP are deeply interwoven. The system does low level analyses before semantic modules are activated, that is tokenization, sentence splitting, multiword creation

from a large lexical database. Then chunking and syntactic constituency parsing which is done using a rule-based recursive transition network: the parser works in a cascaded recursive way to include always higher syntactic structures up to sentence and complex sentence level. These structures are then passed to the first semantic mapping algorithm that looks for subcategorization frames in the lexica made available for English, including VerbNet, FrameNet, WordNet and a proprietor lexicon of some 10K entries, with most frequent verbs, adjectives and nouns, containing also a detailed classification of all grammatical or function words. This mapping is done following LFG principles, where c-structure is turned into f-structure thus obeying uniqueness, completeness and coherence. The output of this mapping is a rich dependency structure, which contains information related also to implicit arguments, i.e. subjects of infinitivals, participials and gerundives. It also has a semantic role associated to each grammatical function, which is used to identify the syntactic head lemma uniquely in the sentence. Finally it takes care of long distance dependencies for relative and interrogative clauses.

When fully coherent and complete predicate argument structures have been built, pronominal binding and anaphora resolution algorithms are fired. Also coreferential processing is activated at the semantic level: they include a centering algorithm for topic instantiation and memorization that we do using a three-place stack containing a Main Topic, a Secondary Topic and a Potential Topic. In order to become a Main Topic, a Potential Topic must be reiterated.

Discourse Level computation is done at propositional level by building a vector of features associated to the main verb of each clause. They include information about tense, aspect, negation, adverbial modifiers, modality. These features are then filtered through a set of rules which have

the task to classify a proposition as either objective/subjective, factual/nonfactual, foreground/background. In addition, every lexical predicate is evaluated with respect to a class of discourse relations. Eventually, discourse structure is built, according to criteria of clause dependency where a clause can be classified either as coordinate or subordinate. We have a set of four different moves to associate to each clause: root, down, level, up. We report here below semantic and discourse structures related to the poem by Sylvia Plath “Edge” shown above.

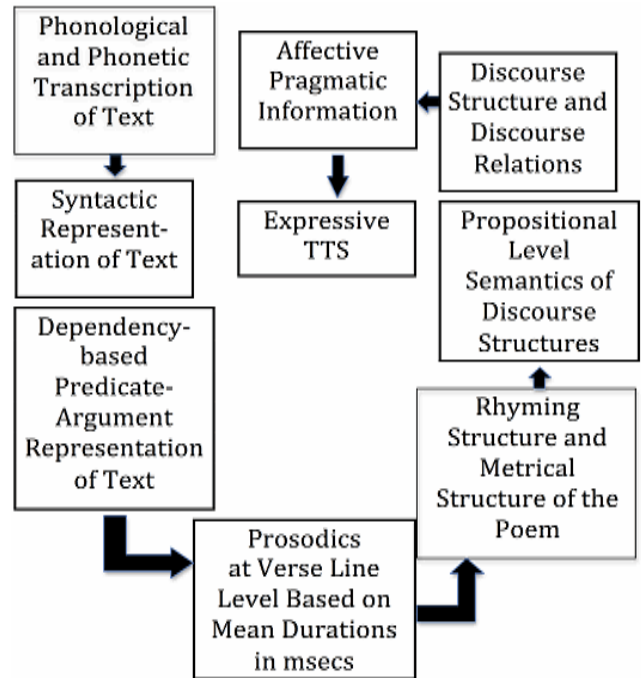


Figure 3. System Architecture Modules for SPARSAR

PROPOSITIONAL SEMANTICS										
Clause No.	Funct/ Role	View	Factivity	Change	Relevance	Aspect	Pred	Tense	Disc_ Rel_	Subject_
39	main/prop	external	factive	null	background	activity	crackle	pres	narration	objective
38	main/prop	external	factive	null	background	activity	drag	pres	narration	objective
31	main/prop	external	factive	culmintd	foregrnd	activity	use	perf	cause	objective
25	xcomp/prop	internal	factive	null	background	activity	moon	pres	narration	objective
24	main/prop	external	factive	null	background	activity	have	pres	narration	objective
23	main/prop	external	factive	culmintd	foregrnd	activity	stare	past	result	objective
18	main/prop	external	factive	culmintd	foregrnd	activity	bleed	past	narration	objective
17	main/prop	external	factive	null	background	activity	fold	perf	narration	objective
16	adj/temp_coinc	external	factive	null	background	activity	stiffen	pres	narration	objective
11	main/prop	internal	factive	null	background	activity	it	pres	narration	objective
10	main/prop	external	factive	culmintd	foregrnd	activity	come	perf	result	objective
7	main/prop	external	factive	null	background	activity	flow	pres	result	objective
6	main/prop	external	factive	culmintd	foregrnd	activity	perfect	perf	result	objective
5	main/prop	external	factive	culmintd	foregrnd	activity	say	past	narration	objective
4	main/prop	external	factive	null	background	activity	seem	pres	narration	objective
3	main/prop	external	factive	null	background	activity	wear	pres	narration	objective
1	xcomp/prop	internal	factive	null	background	state	edge	nil	setting	objective

Figure 4. Propositional semantics for Edge

In Fig.4, clauses governed by a copulative verb like BE report the content of the predication to the subject. The feature CHANGE can either be set to NULL, GRADED or CULMINATED: in this case Graded is not used seen that there are no progressive tenses nor overlapping events.

In the representation of Fig.5, we see topics of discourse as they have been computed by the coreference algorithm, using semantic indices characterized by identifiers starting with ID. Every topic is associated to a label coming from the centering algorithm: in particular, WOMAN which is assigned ID id2 reappears as MAIN topic in clauses marked by no. 15. Also BODY reappears with id7. Every topic is associated to morphological features, semantic inherent features and a semantic role.

DISCOURSE SEMANTICS

Topic Type	Clause No.	Pred	Semant_ Id_	M-Feats Per,Gen,Num	Semantic Inherent Feats	Semantic Role
main	1	edge	id1	[3, neu, sing, [abstrct, legal, nquant, objct],		theme_bound]
poten	3	illusion	id2	[3, nil, nil, [abstrct, inform, danger],		theme_bound]
poten	3	scroll	id3	[3, mas, sing, [abstrct, tecno],		goal]
poten	3	foot	id4	[3, nil, nil, [animat, body_part, objct],		theme_bound]
poten	3	smile	id5	[3, mas, sing, [activ, inform],		goal]
poten	3	toga	id6	[3, nil, nil, [body_part, objct],		theme_bound]
poten	3	dead_body	id7	[3, mas, sing, [objct, hum],		goal]
poten	3	necessity	id8	[3, nil, nil, [place, inform, state],		theme_bound]
poten	3	accomplishment	id10	[3, mas, sing, [abstrct, chang,state],		goal]
main	3	woman	id2	[3, fem, sing, [any, relat, social, hum],		theme]
second	15	garden	id11	[3, neu, plur, [instit, objct, instrum],		agent]
poten	15	child	id12	[3, neu, sing, [any, activ, body_part, objct, relat, social, instrum, hum],		actor]
poten	15	serpent	id13	[3, neu, sing, [animt, objct, instrum],		theme]
poten	15	throat	id14	[3, neu, plur, [body_part, objct, instrum, hum],		loc_origin]
poten	15	stiffen	id16	[3, neu, plur, [instit],		goal]
poten	15	body	id7	[3, neu, sing, [abstrct, activ, body_part, inform, instit, place, objct, instrum, hum],		loc_direct]
poten	15	pitcher	id15	[3, mas, sing, [activ, inform, nquant, objct, relat, social, instrum, hum],		specif]
poten	15	milk	id17	[3, neu, sing, [body_part, edible, objct, hum],		specif]
poten	15	petal	id18	[3, neu, plur, [plant],		agent]
poten	15	flower	id19	[3, neu, sing, [plant, time],		theme]
poten	15	night	id20	[3, neu, sing, [state, time],		specif]
main	21	hood	id21	[3, mas, sing, [objct, instrum, hum],		loc_origin]
poten	21	moon	id22	[3, neu, sing, [event, place, objct, time],		experiencer]
poten	29	sort_of	id23	[3, nil, nil, [abstrct, activ, inform, relat, social, state, tecno, hum],		attr]

Figure 5. Discourse level Semantics for Topic Hierarchy

Eventually, the final computation concerning Discourse Structure is this one:

Sent_ No.	Clause No.	Subject_	Disc_ Rel_	Tense	Pred	Relevance	Disc_ Move	Disc_Struct_ Attach_Node
edge_7	39	objective	narration	pres	crackle	background	level	down(31-39)).
edge_7	38	objective	narration	pres	drag	background	level	down(31-39)).
edge_6	31	objective	cause	perf	use	foreground	up	to(1-31)).
edge_5	25	objective	narration	pres	moon	background	level	down(18-25)).
edge_5	24	objective	narration	pres	have	background	level	down(18-24)).
edge_5	23	objective	narration	pres	stare	foreground	down	down(18-23)).
edge_4	18	objective	narration	past	bleed	foreground	up	to(1-18)).
edge_4	17	objective	narration	perf	fold	background	level	level(11-17)).
edge_4	15	objective	circumstance	pres	stiffen	background	level	level(11-16)).
edge_3	11	objective	narration	pres	it	background	level	level(7-11)).
edge_3	10	objective	narration	pres	come	foreground	level	level(7-10)).
edge_2	7	objective	result	pres	flow	background	down	down(1-7)).
edge_2	5	objective	narration	past	say	foreground	up	to(1-5)).
edge_2	4	objective	narration	pres	seem	background	level	level(1-4)).
edge_2	3	objective	narration	pres	wear	background	level	level(1-3)).
edge_2	6	objective	narration	perf	perfect	foreground	down	down(1-6)).
edge_1	1	objective	setting	nil	edge	background	up	down(nil-1)).

Figure 6. Discourse Semantics for Discourse Structures

Movements in the intonational contours are predicted to take place when FOREGROUND and UP moves are present in the features associated to each clause.

5. Shakespeare's Sonnets Revisited

5.1. Computing Mood from the Sonnets

In the second part of this article we will show data produced by SPARSAR relatively to the relation intervening between Mood, Sound and Meaning in the whole collection of William Shakespeare's Sonnets. This is done to confirm data presented in the sections above, but also and foremost to start to single out those parameters that more effectively and more efficiently can be deemed to be responsible for the popularity and artistic success of a sonnet. As will be made clear from Table 1. below, choice of words by Shakespeare has been carefully done in relating the theme and mood of the sonnet to the sound intended to be produced while reading it. Shakespeare's search for the appropriate word is a well-known and established fact and a statistics of his corpus speak of some 29,000 types, a lot more than any English poet whose corpus has been quantitatively analysed so far [2]. As far as the collection of the sonnets is concerned we have the following figures:

- Total No. of Tokens 18283
- Total No. of Types 3085
- Type/Token Ratio 0.1687
- No. Hapax Legomena 1724
- No. Rare Words 2441

where Rare Words are the union of all Hapax, Trilegomena, Dislegomena and Hapax Legomena. It is rather impressive that the number of Hapax Legomena or Unique words cover more than half the number of Types, precisely 55.58% of the total. This means that there is a unique word every ten tokens. In particular tokens have been computed after separating genitive 's endings and all punctuation. Other interesting information from the rank list, is that 50% of all tokens are covered by the first 67 types, which include the following three separate lists: content words (adjectives, nouns, verbs); then pronouns, possessives and quantifiers; finally conjunction, adverbials and modals, all listed in their descending frequency rank:

- LOVE, BEAUTY, TIME, HEART, SWEET, EYES, FAIR
- MY, I, THY, THOU, ME, THEE, ALL, YOU, IT, HIS, THIS, YOUR, SELF, NO, MINE, THEIR, THEY, HER, HE, THINE

- NOT, BUT, SO, AS, WHEN, OR, THEN, IF, MORE, WILL, SHALL, NOR, YET, THAN, NOW, CAN, SHOULD

If we continue to follow the rank list to include types up to 60% of all tokens, we come up with these three lists:

- MAKE, EYE, TRUE, LIKE, SEE, WORLD, DAY, LIVE, PRAISE, SAY, GIVE, NEW, LIFE, SHOW, TRUTH, DEAR, LOOK, NIGHT, OLD, KNOW, MEN, DEATH, PART, ALONE, BETTER, FACE, FALSE, HEAVEN, ILL, MADE, SUMMER
- ONE, HIM, SHE, THOSE, SOME, SUCH, OWN, EVERY, THESE, THIS, NOTHING
- STILL, WHERE, HOW, THOUGH, MAY, MOST, WELL, WHY, EVEN, SINCE, BEST, THUS, MUST, WOULD, WORTH, BETTER

Now, if we lemmatize EYE/S to one single entry, and sum up their frequency values, $53+40=93$, this would become the second content word after LOVE. It is thus certified that the themes of the sonnets as characterized by main content words are concerned with love, eyes, time, heart and of the beloved partner, who is also sweet and fair. These are the most recurrent content words and constitute main themes: as we will see below, this thematic bias is confirmed by other data, with some exception though.

We assume sonnets can be classified as regards their mood into the following four categories: 1. sonnets with an overall happy mood; 2. sonnets about love with a contrasted mood – the lover has betrayed the poet but he still loves him/her, or the poet is doubtful about his friend's love; 3. sonnets about the ravages of time, the sadness of human condition (but the poet will survive through his verse); 4 sonnets with an overall negative mood. In our first experiment including however only half of the sonnets we came up with the following results:

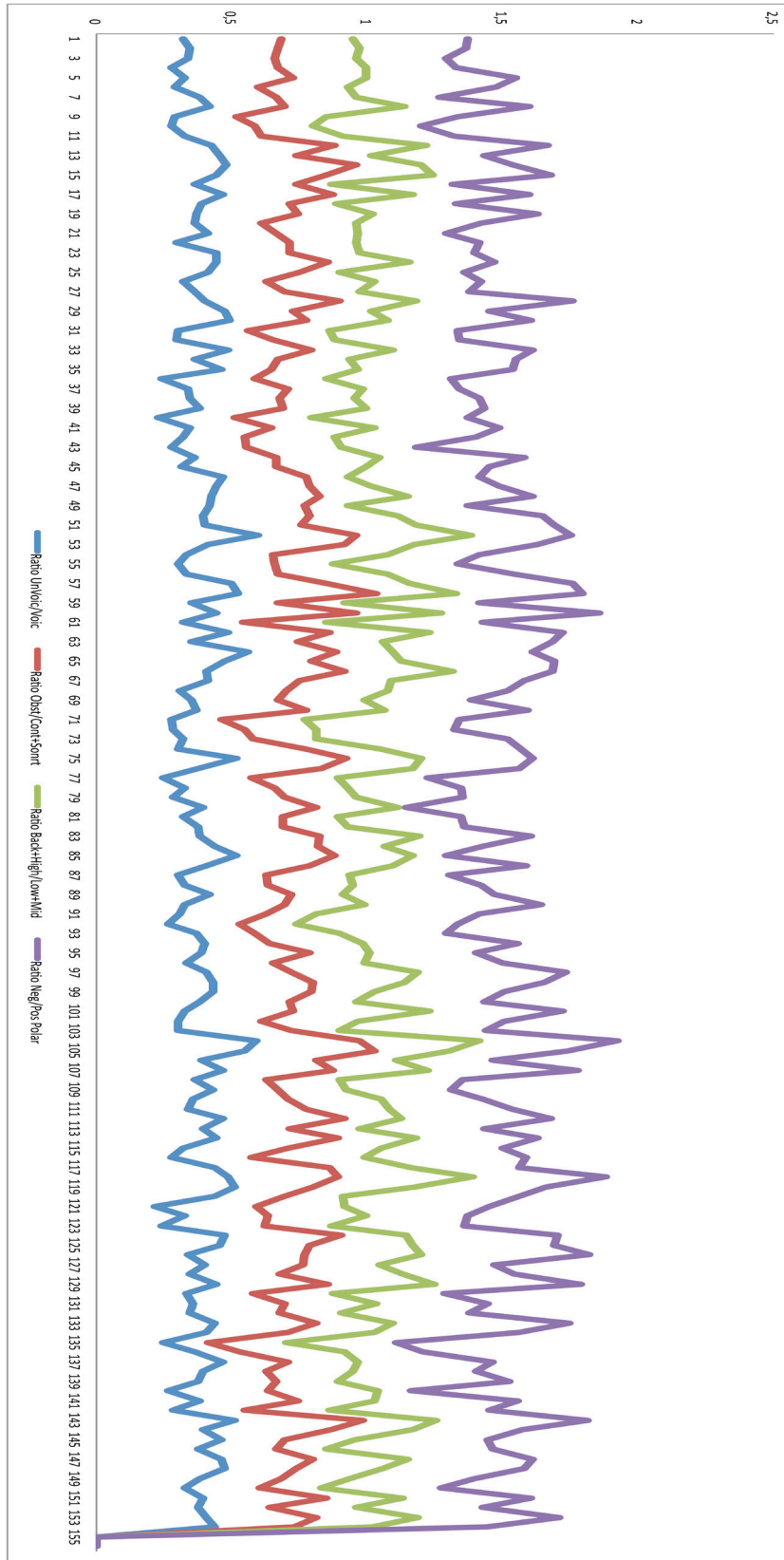
1. POSITIVE peaks (11): sonnet 6, sonnet 7, sonnet 10, sonnet 16, sonnet 18, sonnet 25, sonnet 26, sonnet 36, sonnet 43, sonnet 116, sonnet 130

4. NEGATIVE dips (15): sonnet 5, sonnet 8, sonnet 12, sonnet 14, sonnet 17, sonnet 19, sonnet 28, sonnet 33, sonnet 41, sonnet 48, sonnet 58, sonnet 60, sonnet 63, sonnet 65, sonnet 105

2. POSITIVE-CONTRAST (6): sonnet 22, sonnet 24, sonnet 31, sonnet 49, sonnet 55, sonnet 59

3. NEGATIVE-CONTRAST (1): sonnet 52

Table 1. Comparing Polarity with Sound Properties of Shakespeare's Sonnets: Blue Line = Ratio of Unvoiced/Voiced Consonants; Red Line = Ratio of Obstruents/Continuants+Sonorants; Green Line = Ratio of Marked Vowels/Unmarked Vowels; Violet Line = Ratio of Negative/Positive Polarity Words/Propositions.



Overall, the system has addressed 33 sonnets out of 75 with the appropriate mood selection, 44%. The remaining 42 sonnets have been projected in the intermediate zone from high peaks to low dips. If we look at peaks and dips in Table 1. where all 154 sonnets are considered, and try to connect them to the four possible interpretations of the sonnets, however a new picture comes out. We report the list of high peaks for each class and try intersections between classes. Negatively marked sonnets are the following 39 sonnets:

5, 8, 12, 15, 17, 19, 28, 30, 33, 44, 48, 52, 58, 60, 62, 65, 66, 70, 75, 83, 86, 90, 97, 101, 104, 107, 112, 114, 118, 124, 125, 126, 129, 133, 143, 147, 148, 151, 153

Higher peaks are associated to these 21 sonnets, where 104 is the most negatively marked:

28, 52, 57, 58, 60, 62, 65, 66, 97, 101, **104, 107, 112, 118, 124, 125, 126, 129, 133, 143, 153

High peaks in Back/High vowels is found in the following 59 sonnets:

8, 12, 14, 15, 17, 24, 28, 30, 33, 48, 50, 51, 52, 53, 54, 56, 57, 58, 60, 62, 63, 64, 65, 66, 67, 68, 70, 75, 76, 80, 83, 84, 85, 86, 97, 98, 101, 104, 105, 106, 107, 111, 112, 114, 117, 118, 119, 124, 125, 126, 128, 129, 133, 143, 144, 147, 148, 151, 153

where higher peaks are in this subset made up of 27 sonnets:

12, 14, 15, 17, 24, 28, 48, 51, 52, 53, 57, 58, 60, 62, 66, 75, 76, 83, 85, 97, 101, **104, 105, 107, 118, 126, 143

where again sonnet 104 constitutes the highest peak. As for the remaining 26 sonnets, 18 are included in the list of negative polarity sonnets. If we consider the whole set of negative polarity sonnets compared to the sonnets with high peaks in Back/High vowels the percentage goes up to 89.74% of all sonnets, that is all sonnets are included with the exception of 4 sonnets, 5, 19, 44, 90. Now let's consider the Obstruents, again a class of sound with should go hand in hand with negatively featured sonnets: we have 50 sonnets with high peaks,

12, 14, 15, 17, 24, 28, 30, 33, 46, 47, 48, 49, 50, 51, 52, 53, 57, 58, 60, 62, 64, 66, 75, 76, 80, 83, 84, 85, 86, 95, 98, 99, 104, 105, 106, 107, 111, 112, 114, 117, 118, 119, 124, 129, 133, 143, 144, 147, 151, 153,

where higher peaks are associated to the following 9 sonnets, including 5 exceptionally high ones:

14, 28, 52, 53, **58, **60, **104, **105, **143

these 9 sonnets are all coincident with the ones found in the Back/High sounds. Only six of them however are also found in the Negative Polarity set. If we consider the whole list, the final accuracy match is 28 over 50, that is 56%. However, if we compare Obstruents with Back/High sounds with come up with 45/50 that is 90% accuracy. Another fundamental

class is the one constituted by Unvoiced consonants, which in our hypothesis, should intersect strongly with Negatively marked sonnets and Obstruent consonant sounds. The list if here below, and includes 54 sonnets:

8, 12, 13, 14, 15, 17, 21, 23, 24, 25, 29, 30, 33, 46, 47, 48, 49, 52, 57, 58, 60, 62, 64, 65, 75, 84, 85, 86, 89, 97, 98, 99, 104, 105, 107, 109, 112, 114, 117, 118, 119, 120, 124, 125, 127, 129, 133, 134, 137, 143, 145, 147, 148, 154

Higher peaks are found in the following 18 sonnets:

14, 17, 29, 30, 33, 46, **52, 57, *58, 62, **64, 75, 85, **104, **105, 118, 119, 143

Intersection with Obstruents is 36 over 54 that is 66.67%. Intersection with back/high vowels is 37 over 54 68.52%. Coming now to intersection with Negative Polarity sonnets, only 27 over 54 match, i.e. 50%. However, if we consider only highest peak sonnets, 10 over 18 are found, that is 55.56%.

Eventually, let's consider the opposite class, Positive Polarity classified sonnets. They are the following 27 ones:

3, 4, 7, 9, 10, 11, 16, 18, 21, 36, 43, 55, 71, 72, 77, 80, 85, 87, 93, 108, 109, 122, 123, 130, 135, 140, 150,

with higher peaks in the seven sonnets below, with 4 of them particularly high:

**11, 43, 77, **80, 130, **135, 140

And now we will try to compare these sonnets to the ones characterized by High Low/Mid vowel sounds:

9, 10, 11, 16, 18, 25, 31, 32, 36, 40, 42, 43, 46, 49, 55, 59, 61, 71, 72, 73, 77, 78, 81, 82, 87, 88, 89, 91, 92, 93, 103, 108, 109, 120, 121, 123, 130, 132, 135, 138, 142, 145, 146, 149, 150, 152

with higher dips in the five sonnets below,

**10, 40, 61, **71, 92, 135

Only 19 positively marked sonnets over 27, i.e. 70.37% are also Low/Mid highly marked. Another feature that should compare favourably with Positive sonnets is the one constituted by Continuants/Sonorants, i.e. all those sonnets where the majority of stressed syllables have onset characterized by those consonant sounds. They are the following 38 ones:

6, 9, 10, 11, 20, 21, 26, 31, 32, 36, 40, 42, 43, 54, 55, 56, 59, 61, 71, 72, 73, 77, 78, 91, 92, 93, 102, 108, 109, 116, 121, 122, 123, 130, 135, 136, 142, 150

with higher peaks in the eight sonnets below:

9, 31, 40, 61, **71, 92, **135, 142

Finding intersection with Positively marked sonnets gives the following figure: 18 over 38, i.e. 47.37% a rather low percentage.

Table 2. Raw Statistical Data for Table 1. Showing distributional properties

	Unvoiced	Voiced	Obstruent	Continuant	Sonorant	Low Open	Middle	High Front	Back Closed	Neg-Pol	Pos-Pol
Total	4336	7119	3219	3858	2372	4443	2570	1792	1237	3094	3495
Mean	28,1558	46,2273	20,9026	25,0519	15,4026	28,8506	16,6883	11,6364	8,03247	20,0909	22,6948
St.Dev.	5,2358	8,1742	4,6312	6,1483	4,6323	6,0691	3,783	3,783	3,67	5,0892	4,6839

So eventually, best match is between Obstruent consonant sounds and Back/High vowel sound, 90% match. Then comes the match between Negatively marked sonnets and Back/High vowels, 89.74%. As for Obstruents and Negatively marked sounds, only 56% of them match. The remaining intersections are lower than that, but the important fact is that matching negative features i.e. Unvoiced-Obstruent-Back/High with Positively marked sonnets produces totally bad results:

- only three Unvoiced marked sonnets intersect with Positively marked sonnets
- only two Obstruent marked sonnets intersect with Positively marked sonnets
- the same two Back/High marked sonnets again intersect with Positively marked sonnets

Eventually only two sonnets intersect, 80 and 85 with is also one of the three Unvoiced marked intersecting sonnets, the other two being 21 and 109.

As can be easily gathered from Table 2. data derived from absolute values of phonetic and semantic measurements are very well distributed. Standard Deviations are all in line with their mean. We can notice some differences between best distributed data and worse ones: Polarity data are worse distributed together with Continuants. Best distributed data are Voiced followed by Unvoiced.

5.2. Predicting Popularity from Semantic and Pragmatic Analysis

In this section we will follow approaches carried out in the past by Simonton[55,56] to predict best sonnets from popularity indices and quantitative linguistic analyses. In his papers, he defines his approach to detecting best sonnets as searching for those sonnets that “(a) treat specific themes, (b) display considerable thematic richness in the number of issues discussed, (c) exhibit greater linguistic complexity as gauged by such objective measures as the type-token ratio and adjective-verb quotient, and (d) feature more primary process imagery (using Martindale's Regressive Imagery Dictionary)⁸”. In fact, his approach looks for a coincidence between the outcome of his measures and what is found in popularity indices. These matches were previously elaborated by computer analysis [55] where “the differential popularity of all 154 poems was determined using a 27-item measure that tapped how much and how often a given sonnet was quoted, cited, and anthologized”. As Simonton comments, “the reliability coefficient was commendably high (internal consistency, or Cronbach's, $\alpha = 0.89$), so we can infer the existence of a pervasive consensus on their relative merits”.

Thematic distinction is taken from a list previously elaborated by Huntchins (quoted in Simonton's paper) from the Syntopicon⁹ that served as a detailed topic index to the "Great Books of the Western World". The 154 sonnets were indexed on the basis of some 24 themes.

In his analysis of the sonnets, Simonton mixes up/gauges measurement of agreement with the topic index that is a list based on cultural and semantic-pragmatic criteria, with simple quantitative measures based on the old token-type relation or else the count of unique words, that is words that have been used only once in the whole collection of sonnets. In [55] we find “The type-token ratio also correlates positively with a large number of other variables, including all gauges of thematic richness, unique words, primary process imagery, and the adjective-verb quotient.”(ibid., p.707)

In Simonton[57] we find this association many times, in particular here, where he mixes up “wealth of semantic association” and “greater variety of themes” with “diversity of words”: “Great poems stimulate a wealth of semantic associations. Certainly this would be the case for those sonnets that (a) span a greater variety of themes, (b) use a diversity of words, and (c) favour a highly concrete lexicon over one much more abstract.”(ibid., p.138)

Semantic association to a lexicon that differentiates concrete vs abstract concept is obtained by Simonton using a specific dictionary. “Diversity of words” is related to token-type ratio and unique words ratio.

In particular in [55] we find the following assertion: “The better sonnets are distinguished by a higher type-token ratio, more unique words, a higher adjective-verb quotient, and a more pronounced infusion of primary process imagery with a corresponding dearth of secondary process imagery.”(p.710) A similar statement can be found at pag. 711, pag.713; but also in [56] “As we travel along the continuum from lesser to greater poems, the association between number of words (tokens) and number of different words (types) become ever more pronounced. Hence, the better the sonnet, the more we encounter a poet willing to convert more fully potential variety into actual linguistic variety.”(pag. 262).

As will be shown below, it is not true that unique words and type-token ratio characterize better sonnets. In order to test the correctness of Simonton's quantitative findings relatively to the primary and secondary processes, we will be using the same tool, i.e. Martindale's Dictionary (hence RID) – more on RID below. Something similar has however also been regarded as highly relevant in other approaches (see [44,58]), where quantity of CONCRETE words is used to distinguish good from bad poetry.

As will be clear from our experiment, it is also not true that the most popular sonnets have a majority of concrete or primary process related concepts. We will measure frequency of occurrence of concepts classified by RID using both raw or absolute values, and also treating the entries of the Dictionary in a normalized manner. In addition to the use of the RID, we will be using WordNet [59,60] which requires that in order to make appropriate comparison analyses one has to go from “words” or tokens to lemmata or concepts.

From a more general and linguistically-based point of view, the variety of themes cannot be derived simply by looking at the quantity of unique words present in the text, because it is their semantics that will have to be taken into account. Unique words and/or diversity of words cannot by itself be taken to signify that the text also contains a variety of themes. It is the semantic association of words that has been considered in detail, and this may only ensue from a study based on semantic similarity. Two different words may be unique but be inferentially linked or semantically similar, thus belonging to a same semantic lexical field. Using the RID only indicates a very generic distinction into three psychologically-based set of concepts – primary (sensual, subjective), secondary (rational, abstract, objective), emotions. These three classes are internally further divided up into 65 classes which are highly specific. So a commonality of themes could perhaps be established on the basis of the 65 classes and not simply on the threefold distinction.

Lexical diversity may be important but cannot be the sole criterion by which superior sonnets are selected. So we don't find it very revealing the discovery that there is an increase in the number of unique words in the second quatrain than in other quatrains, and that the sonnets share the tendency for the concluding couples to have rather more words given the number of lines, thus indicating a prevalence of

⁸ freely downloadable from <http://textanalysis.info/pages/category-systems/general-category-systems/regressive-imagery-dictionary.php>

⁹ The Syntopicon is an Index to Great Ideas by Mortimer J. Adler that tries to cover all the ideas that have been created, published and discussed in the western world under the cover terms of some 3000 topic terms. These terms are parceled among 102 ideas - this can be found at <http://www.thegreatideas.org/syntopicon.html>.

monosyllabic rather than polysyllabic words. The reason for that is simply that WORDS are not LEMMATA, and that the latter need to be gauged by their semantic properties. And only in case they don't have any direct inferential relation to the rest of the sonnet's content the UNIQUE words may become important. However, by simply counting different and unique words is not enough to determine deep semantic relations, which alone can detect thematically related word-sense associations.

We also do not find that "As one progresses from the 1st through the 154th sonnet, thematic richness and the type-token ratio decline, while there is a small rise in the frequency of broken lines." [55:707]. As will be made clear in the tables below, both type-token ratio and unique-words ratio remain comparatively high even in the final sonnets. They will appear in Table 4 and 5 below.

As to the use of the Syntopicon, we may notice that the classes favoured by cross-comparison with popularity indices simply discards most popular sonnets by nowadays metrics. By a simple lookup into the web, more popular sonnets have to include the following 25 ones:

1 2 18 23 29 30 33 55 57 64 65 73 75 80 104 109 116 126
129 130 133 138 141 142 147 110.

If we consider sonnets repeatedly mentioned in more than one website, we come up with the following short list made up of 11:

18 29 30 33 73 104 116 126 129 130 138.

Coming now to the list derived from Simonton(1989) system where the author simply matches his popularity indices against the classification based on the Syntopicon, we come up with the following list of 17 sonnets:

1 19 14 15 25 49 55 59 60 63 64 65 81 115 116 123 126

This list is clearly deficient in that it is missing sonnets like 18, 29, 30, 73, 129 and 130 which are by far the most well-known and popular of all. A more coherent list will thus include a union of all these and come up with 34 sonnets, which we will adopt for our investigation in the possibility of extracting popularity indices from our data:

1 2 14 15 18 19 23 25 29 30 33 49 55 59 60 63 64 65 73 75
81 104 109 115 116 123 126 129 130 133 138 141 142 147

¹⁰ we look-up the first two pages of candidates from a Google search engine with keywords "most popular sonnets Shakespeare". The results are taken from the following links:

<http://www.nosweatshakespeare.com/sonnets/shakespeare-famous-sonnets/>
<http://www.enkivillage.com/famous-shakespeare-sonnets.html>
<http://www.william-shakespeare.info/william-shakespeare-sonnets.htm>
<http://www.stagemilk.com/best-shakespeare-sonnets/>
http://www.thehypertexts.com/Masters_of_English_Poetry_and_Lit_erature.html
<http://www.shakespeare-online.com/sonnets/>
https://en.wikipedia.org/wiki/Sonnet_138
<http://bardfilm.blogspot.it/2010/05/seven-best-uses-of-shakespeare-sonnets.html>
<http://www.telegraph.co.uk/books/what-to-read/top-10-romantic-poesms/>
http://www.love-poems.me.uk/shakespeare_william_sonnets.htm
<http://www.usingenglish.com/forum/threads/82317-Which-sonnets-by-Shakespeare-are-the-most-famous>

5.2.1. Themes in the Sonnets: the RID and the Synopticon

Main themes of the sonnets are well-known: from 1 to 126 they are stories about a handsome young man, or rival poet; from 127 to 152 the sonnets concern a mysterious "dark" lady the poet and his companion love. The last two poems are adaptations from classical Greek poems. In the first sequence the poet tries to convince his companion to marry and have children who will ensure immortality. Else love, the poem and poetry will "defeat" death. In the second sequence, both the poet and his companion have become obsessed with the dark lady, the lexicon used is sensual and the tone distressing. These themes are at their highest in the best sonnets indicated above. So we would expect these sonnets to exhibit properties related to popularity that set them apart from the rest.

Now let's consider the list of themes-sonnets organized on the basis of the Synopticon, and match it with the Frequency Rank list of all sonnets – they don't seem to coincide at all. Themes from the Synopticon include the following: Change, Honor, Immortality, Love, Time. In Simonton[40] they are so characterized: "Looking at the thematic measures first, the most successful sonnets are most likely to discuss "the love and hatred of change" (Change 12b), "honor or Tame as a mode of immortality" (Honor 2d), and "immortality through offspring " (Immortality 6b), with less consistent propensities for "the intensity and power of love" (Love 1e), "friendly, tender, or altruistic love " (Love 2b), and "the temporal course of the passions " (Time 7)."(p.710)

Looking up the Frequency Rank list we find that Change appears at rank 199, with frequency 12. However, if we combine it with the word Changes - rank 1597 frequency 1 -, we may consider this theme as ranked 10 positions higher, that is 185, still not in pole position. Honour appears at rank 324 with frequency 7, Immortality does not appear but Immortal is ranked 2153 with frequency 1. Love and Time have already been discussed above and are part of the higher list of types. So perhaps, it is the combination of these themes that makes a sonnet famous or popular not just the presence of one or the other of the main keywords.

We decided to look into the "themes" matter more deeply and discovered that the Immortality theme is in fact present through the lexical field constituted by the keyword DEATH. We thus collected all words related to this main keyword and they are the following ones, omitting all derivations, i.e. plurals for nouns, third person, past tense and gerundive forms for verbs:

BURY, DEAD, DEATH, DECEASE, DECAY, DIE, DISGRACE, DOOM, ENTOMBED, GRAVE, GRIEF, GRIEVANCE, GRIEVE, SCYTHE, SEPULCHRE, TOMB, WASTE

Which we connected to SAD, SADNESS, UNHAPPYNESS, WRINKLE. We ended up by counting 64 sonnets containing this lexical field which can be safely regarded as the most frequent theme of all. We then looked for the opposite meanings, the ones related to LIFE, HAPPY, HAPPYNESS, PLEASURE, PLEASE, MEMORY,

POSTERITY, ETERNITY. In this case, 28 sonnets are the ones mentioning these themes. So, overall, we individuated 92 sonnets addressing emotionally related strong themes. When we combine the two contrasting themes, Death/Eternity, Sadness/Memory, we come up with the following 19 sonnets:

1, 3, 6, 8, 15, 16, 25, 28, 32, 43, 48, 55, 63, 77, 81, 92, 97, 128, 147

This list includes 5 sonnets from the list organized on the basis of the Synopticon categories (1, 15, 25, 55, 63, 81) and two sonnets from the web list (55, 147). There is again only a slight coincidence with the “popular” sonnets. Even if we

search in the 64 sonnets with DEATH themes the intersection is very slight: 10 sonnets coincide with the Synopticon list, and 8 with the web list.

Let’s go back to the RID and its internal organization. RID has some 3200 so-called search patterns, which is roots and words as entries and they are so divided: 1800 belong to primary concepts, 728 to secondary concepts, 616 to emotions. We normalize the absolute values of RID concepts found in every sonnet by dividing up each figure by the total number of entries for that category. In a table below we report results for the eight most popular sonnets, both in absolute value and in normalized ratios.

Table 3. RID absolute and normalized values for the 30 most popular sonnets

Sonnet/RID	Prim_ratio	Sec_ratio	Emot_ratio	PrimA	Sec A	Emot A	MdstP	MdstS	MdstE
sonnet147	0,5746	2,2066	0,8623	10	15	5	-0,3329	1,1981	-0,1811
sonnet142	0,5516	1,2357	2,0696	8	10	7	-0,6771	-0,0976	0,4122
sonnet141	0,9945	0,8487	1,194	15	5	6	0,5274	-1,3933	0,1156
sonnet138	0,3978	2,8856	0,597	6	17	3	-1,0212	1,71646	-0,7744
sonnet133	0,556	1,566	1,669	10	11	10	-0,333	0,1615	1,3022
sonnet130	1,1647	0,9542	0,5593	25	8	4	2,2482	-0,6159	-0,4777
sonnet129	0,5267	1,8388	1,4372	11	15	10	-0,1609	1,1982	1,3022
sonnet126	0,5996	2,3026	0,6749	8	12	3	-0,6771	0,4207	-0,7744
sonnet123	0,5746	2,6153	0,3833	9	16	2	-0,505	1,4573	-1,071
sonnet116	0,7023	1,6345	1,1498	11	10	6	-0,1609	-0,0976	0,1156
sonnet115	0,5223	2,4072	0,7839	10	18	5	-0,333	1,9756	-0,1811
sonnet109	0,6268	2,006	0,9407	8	10	4	-0,6771	-0,0976	-0,4777
sonnet104	1,2312	0,6304	0,7391	25	5	5	2,2481	-1,3934	-0,1811
sonnet81	1,221	1,2872	0	17	7	0	0,8715	-0,875	-1,6643
sonnet73	1,134	1,1614	0,4084	25	10	3	2,2481	-0,0976	-0,7744
sonnet65	1,2233	0,7118	0,6676	22	5	4	1,7319	-1,3934	-0,4777
sonnet64	0,9317	1,4313	0,6992	20	12	5	1,3877	0,4207	-0,1811
sonnet63	0,8897	1,4236	0,8345	16	10	5	0,6995	-0,0976	-0,1811
sonnet60	0,6895	1,7653	1,0348	10	10	5	-0,333	-0,0976	-0,1811
sonnet59	0,8112	2,0768	0,3043	8	8	1	-0,6771	-0,6159	-1,3677
sonnet55	0,9768	1,324	0,6899	17	9	4	0,8715	-0,3567	-0,4777
sonnet49	0,5746	1,7653	1,3798	7	27	4	-0,8492	4,308	-0,4777
sonnet33	0,5028	2,0227	1,2935	23	6	3	1,904	-1,1342	-0,7744
sonnet30	0,5746	1,2259	2,0121	7	14	16	-0,8492	0,939	3,0821
sonnet29	0,7541	1,655	0,9701	12	10	14	0,0112	-0,0976	2,4888
sonnet18	1,1492	1,2036	0,3136	22	9	2	1,7319	-0,3567	-1,071
sonnet15	0,7584	1,7653	0,8278	11	10	4	-0,1609	-0,0976	-0,4777
sonnet14	0,4826	2,8245	0,4139	7	16	2	-0,8492	1,4573	-1,071
sonnet2	1,0671	1,4711	0,2464	26	14	2	0,6995	-1,3934	-0,4777
sonnet1	1,1032	0,8826	0,8278	16	5	4	2,4202	0,939	-1,071

As can be easily gathered, normalized values highlighted show a majority of secondary concepts; whereas in the absolute count on the right hand side, we see the supremacy of primary concepts - 17 over 13. We added three more columns where we compute deviation from the Mean divided up by Standard Deviation for that class. We highlighted then all outliers which are

distributed amongst the three classes, but certainly are in majority found in the Primary column.

These findings are supported by another count which has been done for all sonnets, using semantic lexica like WordNet and other similar tools. In this case, the sonnets are by far characterized by abstract concepts rather than concrete ones.

Table 4. Semantic Classification of best sonnets by categories derived from WordNet

Sonnet/Categs	Abstract	Concrete	Eventive	State	Ratio Neg/Pos	tok-type_rat	tok-rarew_rat
sonnet147	26	1	16	9	0,4651	0,6124	0,1395
sonnet142	28	6	15	12	0,5869	0,5797	0,123
sonnet141	31	4	20	9	0,5306	0,635	0,0949
sonnet138	29	0	16	13	0,4524	0,5946	0,1176
sonnet133	19	4	14	4	0,65	0,542	0,146
sonnet130	24	13	14	4	0,4107	0,5878	0,0743
sonnet129	31	2	18	11	0,537	0,562	0,073
sonnet126	18	5	13	6	0,625	0,661	0,1695
sonnet123	18	6	15	3	0,4872	0,6449	0,0797
sonnet116	21	4	18	3	0,5957	0,6457	0,1575
sonnet115	27	4	20	7	0,4474	0,6449	0,1667
sonnet109	22	5	12	8	0,3846	0,5693	0,1095
sonnet104	24	3	15	6	0,5102	0,662	0,0915
sonnet81	17	10	15	1	0,4545	0,5956	0,169
sonnet75	20	3	10	10	0,4118	0,6087	0,1014
sonnet73	9	6	14	2	0,7105	0,6027	0,1232
sonnet65	17	13	12	5	0,5682	0,6934	0,2117
sonnet64	34	8	23	7	0,5111	0,6412	0,1221
sonnet63	23	8	17	2	0,625	0,6875	0,2031
sonnet60	23	6	18	5	0,5834	0,6947	0,1145
sonnet59	18	6	11	5	0,5	0,687	0,1069
sonnet55	22	12	18	4	0,46	0,6385	0,1308
sonnet49	27	4	19	8	0,4419	0,6029	0,1029
sonnet33	28	11	14	7	0,5128	0,697	0,0985
sonnet30	38	0	30	8	0,534	0,6739	0,1594
sonnet29	25	3	18	6	0,4419	0,6428	0,1
sonnet25	22	5	15	7	0,4634	0,6905	0,1112
sonnet23	31	4	24	6	0,4255	0,6323	0,1691
sonnet19	14	8	13	4	0,6087	0,6838	0,1544
sonnet18	19	7	18	1	0,4419	0,6397	0,1838
sonnet15	27	4	23	4	0,439	0,6846	0,1077
sonnet14	31	6	18	11	0,3415	0,637	0,1037
sonnet2	30	7	23	5	0,3953	0,6594	0,1739
sonnet1	18	4	10	7	0,4242	0,6589	0,2016

As can be noticed from this table, neither tok-unique ratio nor type-token ratio show a consistent reduction in the higher section of the list. On the contrary, in Table 5 we can see that tok-unique list has high peaks even in a sonnet like 147. As to Abstract and Concrete categories ones in Table 4. their supremacy is total. In their new compounded version where we included event and states, the supremacy of Abstract is still remarkable even though not as clear as in the previous separate classification.

Table 5. Semantic Categories from WordNet collapsed to resemble the ones in RID

Sonnet/Categs	Abstract+State	Concrete+Event	Ratio Neg/Pos	Ratio Tok-Uniq
sonnet147	35	17	0,4651	0.1085
sonnet142	40	21	0,5869	0.1014
sonnet141	40	24	0,5306	0.0876
sonnet138	42	16	0,4524	0.0809
sonnet133	23	18	0,65	0.0764
sonnet130	28	27	0.4107	0.0541
sonnet129	42	20	0,537	0.0657
sonnet126	24	18	0.625	0.0847
sonnet123	21	21	0,4872	0.0869
sonnet116	24	22	0.5957	0.0866
sonnet115	34	24	0,4474	0.1087
sonnet109	30	17	0,3846	0.073
sonnet104	30	18	0,5102	0.0704
sonnet81	18	25	0,4545	0.0493
sonnet75	30	13	0,4118	0.0869
sonnet73	11	20	0.7105	0.0958
sonnet65	22	25	0.5682	0.1095
sonnet64	41	31	0,5111	0.084
sonnet63	25	25	0,625	0.0625
sonnet60	28	24	0,5834	0.1221
sonnet59	23	17	0,5	0.084
sonnet55	26	30	0.46	0.1154
sonnet49	35	23	0,4419	0.0809
sonnet33	35	25	0,5128	0.1364
sonnet30	46	30	0.534	0.1087
sonnet29	31	21	0.4419	0.1
sonnet25	29	20	0,4634	0.0794
sonnet23	37	28	0,4255	0.0662
sonnet19	18	21	0,6087	0.0882
sonnet18	20	25	0.4419	0.0588
sonnet15	31	27	0,439	0.1
sonnet14	42	24	0,3415	0.0889
sonnet2	35	30	0,3953	0.1014
sonnet1	25	14	0,4242	0.062

When we consider semantic classes from WordNet we see that almost all sonnets have a clear majority of Abstract concepts. However, if we group together Eventive and Concrete thus resembling what has been done in RID with Primary Concepts, and Abstract with States, again what has been done in RID with Secondary Concepts we see that three sonnets have a majority of Concrete+Eventive categories, while the others still show the same distribution. In both tables we added a column containing the ratio of Negative Polarity items in each sonnet, divided up by the sum of both Negative and Positive Polarity items. Thus, if the ratio is below 0.5 it means that Positive Polarity items are the majority, otherwise the opposite applies. In particular we see then, that sonnet73, sonnet126 and sonnet116 are characterized by a majority of negative polarity concepts. On the contrary, sonnet130, sonnet18, sonnet29 and sonnet55 have a majority of positive polarity items.

If we try to match RID categories with this classification, we see that sonnets 29 and 30 having the majority of Abstract and Abstract+State are also characterized as having the majority of absolute values for Emotions. Sonnets 55, Sonnet 65 and 73 have the majority of Concrete+Event category and are so characterized in the Primary absolute classification. As for the remaining sonnets, they receive opposite classification in the two systems: they are mostly regarded as belonging to Primary concepts in absolute values, but they are shifted to Secondary concepts in the normalized measure, with the exception of sonnet 130.

In Table 5 we reported values for Token/Unique Words ratio for the sonnets listed. We can easily see that sonnet 130 is the one with the least number of Unique words. This does not mean that later sonnets suffer from lack of inspiration or search for the right word as for the first ones: sonnet 154 has a ratio of 0.1338 which is much higher than the one associated to sonnet 1, 0.06201. However from sonnet 130 onward the ratio is always around 0.05, up to sonnet 152 which has a ratio of 0.06896.

To give a comparison gauge of how best sonnets are ranked in respect to the rest of the collection, here below we show the list of highest peaks in the Primary, Secondary and Emotion classification.

PRIMARY CONCEPTS: SONNETS WITH HIGHEST PEAKS

2, 43, 46, **73**, 82, 104, **130**, 137, 153

PRIMARY CONCEPTS: SONNETS WITH HIGH PEAKS

7, 34, 54, **65**, 81, 94, 104, 128, 154

SECONDARY CONCEPTS: SONNETS WITH HIGHEST PEAKS

49, 57, 76, 115

SECONDARY CONCEPTS: SONNETS WITH HIGH PEAKS

4, 14, 27, 33, 39, 50, 57, 62, 67, 76, 79, 85, 89, 123, 138, 149

EMOTIONS CONCEPTS: SONNETS WITH HIGHEST PEAKS

8, 25, **30**, 40, 92

EMOTIONS CONCEPTS: SONNETS WITH HIGH PEAKS

8, 26, 31, 37, 41, 43, 70, 91, 131, 142

As can be seen, sonnets 18, 29, 55, 116, 126 don't figure in any of the lists above. We only have sonnets 73 and 130 characterized by highest peaks in Primary Concepts, and sonnet 30 characterized by highest peak in Emotions Concepts. So Simonton's wrongly assumed prediction that most popular concepts should be very high in Primary concepts is not borne out by the data¹¹. High presence of Unique words in the whole poem or in the final couplets is also not borne out by the data, with the only exception of sonnet 18.

Table 6. Correlation data for three classes from RID both in absolute and normalized values

Correl/Data_typ	Primary/Second	Second/Emotions	Primary/Emotions
Absolute values	-0,271997406	-0,010754049	-0,324919596
Normalized values	-0,705408609	-0,137571741	-0,604917352
Weighted values	-0,252227071	-0,000696759	-0,295436879

Here above in Table 6. we report data showing correlation between the three main classes, both for absolute and normalized data. Best results are obtained for normalized values for Primary and the two other classes, with primary/secondary relation best. However, none of the data reach a reasonable level of correlation. In the figure below I show the corresponding data.

5.3. Using Semantic and Syntactic Data for Readability Evaluation

As a last evaluation we will be using indices derived from SPARSAR for readability[61] checking. These indices are computed on the basis of quantitative counts for modifier heaviness, number of constituents per clause, number of clauses per sentence, number of inflected verbs vs. uninflected ones, presence of negation at propositional level, as well as quantity of abstract and negative words. The position in the graded scale is determined by what we called *A General Description map* which includes seven Macro Indices with a statistical evaluation of such descriptors as: Semantic Density; General Poetic Devices; General Rhetoric Devices; Prosodic Distribution; Rhyming Schemes; Metrical Structure, which we already presented above.

¹¹ D.K.Simonton kindly sent me his data for the popularity indexing, which correlate better with Primary Classification than with Secondary Classification (0.240 vs. 0.193). I have two criticisms about these results: first of all the correlation coefficients are fairly low; secondly, the fact that some correlation exists between popularity indices and Primary Classification simply indicates that the two series have some similarities from the point of view of their distribution. If we look at the Standard Deviations we see however that Popularity data have a St.Dev. of 7.8276599581, with a Mean of 5.1234; while Primary data St.Dev. is 2.121 with a Mean of 11.63. So there is no similarity at all between the two distributions and the Pearson correlation coefficient only indicates a vague correlation.

Table 7. General Semantic and Poetic/Prosodic Indices

Sonnet/ Categs	Poetic Devices	Semantic Density	Prosodic Str.Distr	Deep Concept. Index	General Sem.Index
sonnet147	1330	278.517	314.395	133	5.472
sonnet142	11440	253.869	467.017	1144	8.94
sonnet141	8120	273.887	374.358	812	6.121
sonnet138	6330	334.349	377.003	633	6.479
sonnet133	9330	278.627	343.427	933	1.983
sonnet130	7110	166.201	329.918	711	5.25
sonnet129	3750	278.82	270.721	375	4.583
sonnet126	1530	298.157	440.098	153	7.753
sonnet123	2750	279.572	461.02	275	5.092
sonnet116	1380	286.597	509.082	138	6.516
sonnet115	1790	269.08	550.549	179	7.31
sonnet109	2710	262.124	411.295	271	8.015
sonnet104	6240	77.239	769.653	624	3.658
sonnet81	5190	57.596	379.246	519	10.528
sonnet75	4410	61.227	424.205	441	3.482
sonnet73	7530	65.504	453.036	753	3.397
sonnet65	6960	91.612	498.608	696	9.208
sonnet64	5460	70.24	353.604	546	3.599
sonnet63	5580	84.47	271.008	558	5.889
sonnet60	3930	69.295	341.691	393	1.773
sonnet59	3870	77.438	352.702	387	5.645
sonnet55	4800	132.169	317.378	480	3.437
sonnet49	2110	55.107	394.634	211	3.549
sonnet33	1450	156.494	361.525	145	2.983
sonnet30	6080	92.902	361.024	608	1.943
sonnet29	2340	41.881	393.456	234	10.067
sonnet25	1430	68.802	294.442	143	4.627
sonnet23	1880	71.537	258.601	188	4.844
sonnet19	1940	99.525	552.386	194	1.94
sonnet18	2210	116.104	351.309	221	3.738
sonnet15	1520	73.686	494.557	152	4.542
sonnet14	1780	113.237	427.438	178	7.814
sonnet2	7290	77.544	336.744	729	7.699
sonnet1	1680	76.641	297.211	168	3.738

Semantic Density classification for the 34 sonnets is the index that encodes syntactic and semantic complexity at clause and sentence level. As can be easily seen, the highest density is associated with sonnets 126 and 116, with sonnet 29 having the lowest value.

In figure 7. below, we show a Graded Evaluation Scale of all Sonnets limited though to the first half of the picture where the best sonnets are situated. The evaluation is done on the basis of parameters listed in Table 7. and then further graded on the respective position of each sonnet in each list.

On a first evaluation, we considered as positive two parameters which are critical, Semantic Density and Deep Conceptual Index. Accordingly, best sonnets are the following 41 sonnets in that order:

Sonnet18, Sonnet40, Sonnet42, Sonnet29, Sonnet87, Sonnet51, Sonnet146, Sonnet38, Sonnet4, Sonnet46, Sonnet 37, Sonnet12, Sonnet26, Sonnet27, Sonnet144, Sonnet19, Sonnet74, Sonnet2, Sonnet150, Sonnet122, Sonnet68, Sonnet29, Sonnet72, Sonnet137, Sonnet52, Sonnet152, Sonnet116, Sonnet28, Sonnet100, Sonnet141, Sonnet48, Sonnet43, Sonnet91, Sonnet134, Sonnet58, Sonnet78, Sonnet133, Sonnet69, Sonnet66, Sonnet76, Sonnet109, Sonnet104.

In this way, we only manage to include 10 from the popularity list of 34 above. We then reversed the grading for the two classes indicated above, that is, Semantic Density and Deep Conceptual Indices, regarding them as negatively contributing to the grading. In this way, those sonnets with higher values will be graded lower and will appear at the end of the list. The results are visible in the following Figure 7.

General Graded Evaluation Scale

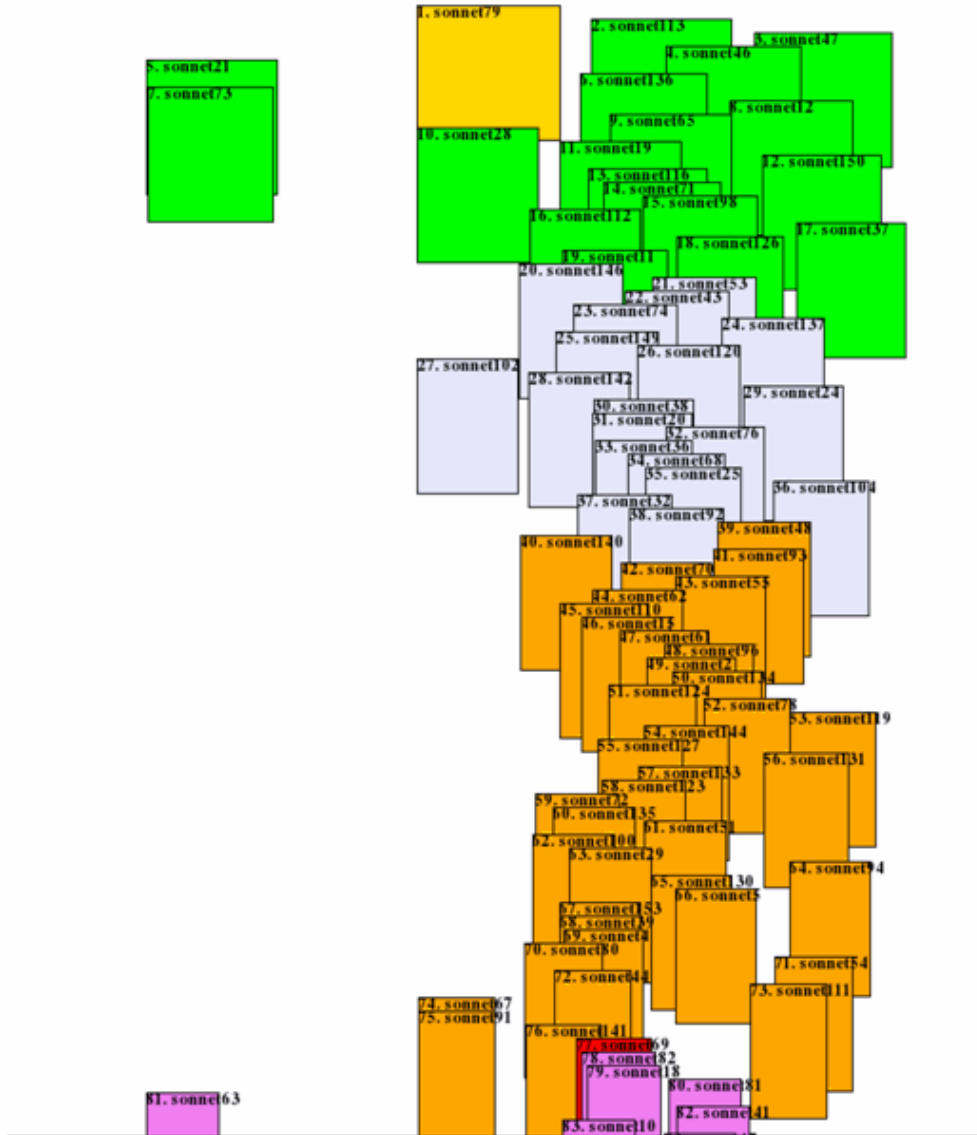


Figure 7. Graded Scale for first half of sonnets contained in the List with rewards reversed for difficult parameters

We considered again half of the list with the best in the order indicated by the new graded evaluation. In this case we correctly addressed 13 sonnets from the 23 of the web lists (56.52%), which we marked with a star below,

*18 23 *29 30 33 *55 57 64 *65 *73 75 *80 *104 109
*116 *126 129 *130 *133 138 *141 *142 147

seven from the most mentioned sonnets on the web (63.64%), again marked by a star below,

*18 *29 30 33 *73 *104 *116 *126 129 *130 138

and 10 from the Synopticon list (58.82%),

1 14 *15 *19 *25 *49 *55 59 60 63 64 *65 *81 115 *116
*123 *126.

6. Conclusions

From the data reported above, it is hard to understand what criteria would be best choice for the individuation of most popular sonnets. It seems clear, however, that neither themes nor readability indices are sufficient by themselves to identify them all. Nor do evaluations based on semantic/pragmatic criteria derived from existing lexica help in the final classification. We surmise that an evaluation of how much popular a poem can be should also take into account cultural issues which have not been tackled by this study. Also a better way to assess the intertwined contribution of rhetoric, poetic and prosodic devices could lead to a better understanding of the value of each sonnet. This is however hard to produce automatically with current technologies. In particular, the contribution of rhetoric devices, like similes and metaphors, is hard to compute consistently for all sonnets: Shakespeare's best virtue was his subtlety in generating a great quantity of secondary meanings from simple juxtaposition of terms and images.

So eventually, what SPARSAR can do is help practitioners in that direction without giving a final complete result, but leave the user to combine different schemes, graphs, tables and other data together in the puzzle constituted by poetry that aims at excellence and lasts forever, like the one we have been commenting in this article. Some of the results are certainly successful in coming to terms with the hidden architecture of the poems: and this is the case for the sound-mood and colour connection.

REFERENCES

- [1] Delmonte R., 2013a. Transposing Meaning into Immanence: The Poetry of Francis Webb, in *Rivista di Studi Italiani*, Vol. XXXI, n° 1, 835-892.
- [2] Delmonte R. 2013b. Computing Poetry Style, in C. Battaglino, C. Bosco, E. Cambria, R. Damiano, V. Patti, P. Rosso (eds.), *Proceeding ESSEM - Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI (ESSEM 2013)*, CEUR Workshop Proceedings, Torino, 148-155, <http://ceur-ws.org/Vol-1096/>.
- [3] Delmonte R. 2014. A Computational Approach to Poetic Structure, Rhythm and Rhyme, in R. Basili, A. Lenci, B. Magnini (eds), *Proceedings of CLiC-it - The First Italian Conference on Computational Linguistics*, Pisa University Press, Vol.1, 144-150.
- [4] Delmonte R. & C. Bacalu. 2013. SPARSAR: a System for Poetry Automatic Rhythm and Style AnalyzeR, *SLATE 2013 - Demonstration Track*, Grenoble.
- [5] Delmonte R. & A.M. Prati. 2014. SPARSAR: An Expressive Poetry Reader, *Proceedings of the Demonstrations at the 14th Conference of the EACL*, Gotheborg, 73-76.
- [6] Delmonte R., 2015. SPARSAR - Expressivity in TTS and its Relations to Semantics, *Invited Talk at AISV 2015*, Bologna.
- [7] Jakobson, R. 1978. *Six lectures on sound and meaning* (Trans.: J. Mepham). Cambridge: MIT Press (Original work published in 1976).
- [8] Jakobson, R., & Waugh, L. 1978. *The sound shape of language*. Bloomington: Indiana University Press.
- [9] Mazzeo, M. 2004. *Les voyelles colorées: Saussure et la synesthésie*. *Cahiers Ferdinand de Saussure*, 57, 129-143.
- [10] Tsur, Reuven. 1992. *What Makes Sound Patterns Expressive: The Poetic Mode of Speech-Perception*. Durham N. C.: Duke UP.
- [11] Fónagy, Iván (1971) "The Functions of Vocal Style", in Seymour Chatman (ed.), *Literary Style: A Symposium*. London: Oxford UP, 159-174.
- [12] Macdermott M.M. 1940. *Vowel Sounds in Poetry: Their Music and Tone Colour*, *Psyche Monographs*, No.13, London: Kegan Paul, 148 pp.
- [13] Mohammad Saif, *Colourful Language: Measuring Word-Colour Associations*, 2011a. In *Proceedings of the ACL 2011 Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, June 2011, Portland, OR.
- [14] Mohammad Saif, *Even the Abstract have Colour: Consensus in Word Colour Associations*, 2011b. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, June 2011, Portland, OR.
- [15] Reddy, Sravana & John Goldsmith, 2010. An MDL-based approach to extracting subword units for grapheme-to-phoneme conversion, in *Proc. HLT-NAACL*, 713-716.
- [16] Genzel Dmitriy, J. Uszkoreit, and F. Och. 2010. "Poetic" statistical machine translation: Rhyme and meter. In *Proceedings of EMNLP*.
- [17] Byrd Roy J. and M. S. Chodorow. 1985. Using an online dictionary to find rhyming words and pronunciations for unknown words. In *Proceedings of the 23rd Annual Meeting of ACL*, 277-283.
- [18] Sonderegger, Morgan. 2011. Applications of graph theory to an English rhyming corpus. *Computer Speech and Language*, 25:655-678.
- [19] Greene E., T. Bodrumlu, K. Knight. 2010. Automatic

- Analysis of Rhythmic Poetry with Applications to Generation and Translation, in *Proceedings of the 2010 Conference on EMNLP*, 524–533.
- [20] Tsur Reuven. 2012. *Poetic Rhythm: Structure and Performance: An Empirical Study in Cognitive Poetics*, Sussex Academic Press, 472.
- [21] Gervas, P. (2001). An expert system for the composition of formal spanish poetry. *Knowledge-Based Systems*, 14(3):181–188.
- [22] Gervas, P. (2010). Engineering linguistic creativity: Bird flight and jet planes. In *Proceedings of the NAACL HLT 2010 Second Workshop on Computational Approaches to Linguistic Creativity*, pages 23–30.
- [23] Toivanen, J. M., Toivonen, H., Valitutti, A. & Gross, O. 2012. Corpus-based generation of content and form in poetry. In *International Conference on Computational Creativity*, 175–179.
- [24] Agirrezabal Manex, Bertol Arrieta, Aitzol Astigarraga, Mans Hulden, 2013. POS-tag based poetry generation with WordNet, *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 162–166.
- [25] Manurung Hisar Maruli, G. Ritchie, and H. Thompson. 2000a. Towards a computational model of poetry generation. In *Proceedings of AISB Symposium on Creative and Cultural Aspects and Applications of AI and Cognitive Science*, 17-20.
- [26] Manurung M.H., G. Ritchie, H. Thompson. 2000b. A Flexible Integrated Architecture For Generating Poetic Texts. in *Proceedings of the Fourth Symposium on Natural Language Processing (SNLP 2000)*, Chiang Mai, Thailand, 7-22.
- [27] Reddy, Sravana & Kevin Knight. 2011. Unsupervised Discovery of Rhyme Schemes, in *Proceedings of the 49th Annual Meeting of ACL: shortpapers*, 77-82.
- [28] Baayen R. H., R. Piepenbrock, and L. Gulikers. 1995. *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium.
- [29] Hartman, C. (2005). *The Scandroid Manual*. Online + Hartman, C. (2004). *Charles Hartman Programs*. Online
- [30] Heuser, R. (2015). *Stanford Literary Lab Github Account*. Online
- [31] Algee-Hewitt, M., Heuser, R. Kraxenberger, M., Porter, J., Sensenbaugh, J., and Tackett, J. (2014). *The Stanford Literary Lab Transhistorical Poetry Project Phase II: Metrical Form*. *Proceedings, Stanford University, Lausanne*.
- [32] Plamondon M. R.: Virtual verse analysis: Analysing patterns in poetry. *Literary and Linguistic Computing* 21, suppl 1 (2006), 127–141.
- [33] Manish Chaturvedi, Gerald Gannod, Laura Mandell, Helen Armstrong, Eric Hodgson. 2012. *Myopia: A Visualization Tool in Support of Close Reading*. *Digital Humanities* 2012.
- [34] M. Agirrezabal, B. Arrieta, A. Astigarraga, M. Hulden. 2013. "ZeuScansion: a tool for scansion of English poetry," *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing*.
- [35] Nina McCurdy, Vivek Srikumar, Miriah Meyer, 2015. *RhymeDesign: A Tool for Analyzing Sonic Devices in Poetry*, *Computational Linguistics for Literature*, ACL, Denver, 12-22.
- [36] Manish Chaturvedi, Gerald Gannod, Laura Mandell, Helen Armstrong, Eric Hodgson. 2012. *Rhyme's Challenge: Hip Hop, Poetry, and Contemporary Rhyming Culture*. Oxford University Press, *Literary Criticism*.
- [37] Karteek Addanki and Dekai Wu. 2013. Unsupervised Rhyme Scheme Identification in Hip Hop Lyrics using Hidden Markov Models. *Proceedings of the 1st International Conference on Statistical Language and Speech Processing (SLSP - 2013)*, Tarragona, Spain.
- [38] Hussein Hirjee and Daniel Brown. 2009. Automatic Detection of Internal and Imperfect Rhymes in Rap Lyrics. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*. pages 711-716.
- [39] Susan Bartlett, Grzegorz Kondrak, and Colin Cherry. 2009. On the syllabification of phonemes. In *Proceedings of Human Language Technologies: NAACL '09*. ACL, Stroudsburg, PA, USA, 308316.
- [40] Alfie Abdul-Rahman, Julie Lein, Katharine Coles, Eamonn Maguire, Miriah Meyer, and Martin Wynne, Chris Johnson, Anne E. Trefethen, Min Chen. 2013. Rule-based Visual Mappings - with a Case Study on Poetry Visualization. In *Computer Graphics Forum*, 32(3):381-390.
- [41] Keim D. A., Oelke D.: Literature fingerprinting: A new method for visual literary analysis. In *IEEE VAST (2007)*, pp. 115–122.
- [42] Oelke D., Bak P., Keim D., Last M., Danon G.: Visual evaluation of text features for document summarization and analysis. In *IEEE VAST (Oct. 2008)*, pp. 75–82.
- [43] Wattenberg M., Viégas F. B.: The Word Tree, an interactive visual concordance. *IEEE Trans. Visualization & Comp. Graphics* 14, 6 (Nov. 2008), 1221–1228.
- [44] Kao Justine and Dan Jurafsky. 2012. "A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry". in *Proc. NAACL Workshop on Computational Linguistics for Literature*.
- [45] Kaplan, D. (2006). *Computational analysis and visualized comparison of style in American poetry*. Unpublished undergraduate thesis.
- [46] Kaplan, D., & Blei, D. (2007). A computational approach to style in American poetry. In *IEEE Conference on Data Mining*.
- [47] Hayward, M. (1991). A connectionist model of poetic meter. *Poetics*, 20, 303-317.
- [48] Hayward, M. (1996). Application of a connectionist model of poetic meter to problems in generative metrics. *Research in Humanities Computing* 4. (pp. 185-192). Oxford: Clarendon Press.
- [49] Bacalu C., Delmonte R. 1999. Prosodic Modeling for Syllable Structures from the VESD - Venice English Syllable Database, in *Atti 9° Convegno GFS-AIA*, Venezia.
- [50] Bacalu C., Delmonte R. 1999. Prosodic Modeling for Speech Recognition, in *Atti del Workshop AI*IA - "Elaborazione del Linguaggio e Riconoscimento del Parlato"*, IRST Trento, pp.45-55.
- [51] Tsur Reuven. 1997. "Poetic Rhythm: Performance Patterns

- and their Acoustic Correlates". *Versification: An Electronic Journal Devoted to Literary Prosody*. (<http://sizcol1.u-shizuoka-ken.ac.jp/versif/Versification.html>)
- [52] Delmonte R., et al. 2005. VENSES – a Linguistically-Based System for Semantic Evaluation, in J. Quiñero-Candela et al.(eds.), *Machine Learning Challenges*. LNCS, Springer, Berlin, 344-371.
- [53] Delmonte R. and V. Pallotta, 2011. Opinion Mining and Sentiment Analysis Need Text Understanding, in "Advances in Distributed Agent-based Retrieval Tools", Springer, 81-96.
- [54] Delmonte R. 2014. ITGETARUNS A Linguistic Rule-Based System for Pragmatic Text Processing, in C. Bosco, P. Cosi, F. Dell'Orletta, M. Falcone, S. Montemagni, Maria Simi (eds.), *Proceedings of Fourth International Workshop EVALITA*, Pisa University Press, Vol. 2, 64-69.
- [55] Simonton, Dean Keith 1989. Shakespeare's Sonnets: A Case of and for Single-Case Historiometry, *Shakespeare's sonnets: A case of and for single-case historiometry*. *Journal of Personality*, 57, 695-721.
- [56] Simonton, Dean Keith 1990. Lexical choices and aesthetic success: A computer content analysis of 154 Shakespeare sonnets. *Computers and the Humanities*, Vol. 24, No. 4 (Aug., 1990), pp. 251-264.
- [57] Simonton, Dean Keith 2006. Historiometric methods. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 319-335). New York: Cambridge University Press.
- [58] Brysbaert, M., Warriner, A.B., & Kuperman, V. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46, 904-911.
- [59] Miller, George A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* Vol. 38, No. 11: 39-41.
- [60] Christiane Fellbaum (1998, ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- [61] Sara Tonelli, Ke Tran Manh, Emanuele Pianta. 2012. "Making Readability Indices Readable", in *Proceedings NAACL-HLT 2012 Workshop on Predicting and Improving Text Readability for target reader populations*, Montréal, Canada, 40-48.