Virtual Verse Analysis: Analysing Patterns in Poetry

Marc R. Plamondon
Department of English, University of Toronto, Canada

Abstract

This article discusses the problem of computer identification of some basic patterns in poetry: rhythm and rhyme. The program AnalysePoems uses the *Representative Poetry Online* (http://rpo.library.utoronto.ca) corpus of poems to identify these patterns. The goal of the rhythm identification is not to produce a perfect metrical scansion of a poem, but to identify the dominant metre with a reasonable degree of confidence. Twelve example poems are used to show that the program is successful in its ability to identify the dominant metre, if one exists, even when some of the words of the poem are unknown to the computer. The computer is usually able to identify the number of syllables and dominant accent of these previously unknown words and then use this new data in its analysis of more poems. Similarly, an analysis of a poem's rhyme scheme allows the computer to identify rhyming pairs it had not previously encountered and use this new information to analyse subsequent rhyme schemes. The analysis of basic patterns lays the groundwork for the analysis of more complex, less obvious patterns. The changes in rhythmic confidence values from first to final analysis suggest a possible measure of the complexity and beauty of a poem's rhythms.

Correspondence:
Marc R. Plamondon,
130 St George Street,
7th floor, John P. Robarts
Research Library,
University of Toronto,
Toronto,
Ontario,
M5S 1A5.
E-mail:
mplamond@chass.
utoronto.ca

1 Introduction

The *Representative Poetry Online* (RPO) website (http://rpo.library.utoronto.ca) was recently converted from a collection of static HTML files to a relational database of poems and automatically generated HTML files. The impetus behind the conversion was the increasingly cumbersome task of editing title and first line indices every time a new poem was added to the site. Ian Lancashire, the general editor of RPO, approached the Web Development Group of the Information Technology Services of the University of Toronto Libraries with his concern, and they suggested the conversion to a database-centred site.[1] Saving the poems in a relational database allows for the automation of the creation of indices; it also allows for other features, such as the site's contextual word concordance. Subsequent modifications and additions to the site can be implemented without the need to restructure the site or the database. The following article describes a program, AnalysePoems, that automates the editorial task of identifying the dominant metre and rhyme scheme of each poem: the program thus analyses two of the basic types of patterning in poetry. The program uses confidence values to help determine the patterns. With confidence values, the computer is able to handle elements it does not yet recognize and extrapolate data values for these unknown elements. As a result, the collection of data grows automatically, without the need of human input. With data extrapolation, the computer learns to analyse new poems on its own, though periodic data checking is expected to maintain integrity.

Literary and Linguistic Computing, Vol. 21, Suppl Issue, 2006.

doi:10.1093/llc/fql011 Advance Access Published on 3 March 2006

This project is hardly the first attempt to analyse poetic metre with a computer. Some noteworthy attempts are by Logan (1982), Hayward (1991, 1996a, 1996b), and Hartman (1996). This attempt differs from those of these predecessors in a few ways. They almost invariably focus on iambic pentameter lines. For example, Malcolm Hayward says, 'I have chosen to work with English iambic pentameter rhythm as it is the most common verse form in English poetry' (1996b, p. 186). This statement is perhaps more contentious than Hayward recognizes: initial estimates indicate that little more than half of the poetic corpus of RPO is composed of iambic pentameter lines. Perhaps Hayward and others are operating under the assumption that iambic pentameter is the English verse form most often encountered in undergraduate English classrooms; and perhaps they are right. But Ian Lancashire's discussion of the RPO corpus, in 'Free Poetry for the World: Editing *Representative Poetry Online*,' suggests that the corpus blurs the distinction between 'academic' and 'popular' poems (see especially Lancashire, 2002, p. 21). RPO includes many poems in iambic tetrameter and in trochaic catalectic tetrameter; it also includes a substantial number of poems with more complex stanza structures, such as two lines of iambic tetrameter alternating with one line of iambic trimeter. The program discussed here allows for such possibilities.

The goal of other computer analyses of poetic metre is sometimes to see how accurately a computer can approximate a natural reading of a line of poetry. This, of course, is a fascinating endeavour, but beyond the concerns of this article. Hayward's results are intriguing, but every line of iambic pentameter requires sixty elements of input data (six values for each of the ten syllables in the line) beyond a pre-identification of the syllabic divisions in the line. Doing this for over 200,000 lines of poetry in the RPO corpus would be time-consuming, to say the least. The program described here requires only the syllabic division of the words and an identification of the usual or probable syllabic emphases for each word. Indeed, if a word is not already known by the computer, the program can not only compensate for the missing data, but can suggest probable data for the word if the conditions are right.

The goal of AnalysePoems is not to approximate a natural reading of the poem, nor to evaluate the degree of metricality of a line or poem, nor to identify a metric fingerprint used to identify authors or periods of poetic verse.[2] Instead, it is to automate the identification of the dominant metrical pattern of a poem and to describe some basic elements of the structure of the poem, such as the number of syllables per line and whether all lines are of the same syllabic length or whether there are variations in the syllabic lengths of the lines in an identifiable pattern. The results of this automatic identification can then be inserted into the RPO database, saving the general editor from having to identify the dominant metre and rhyme scheme of every poem.

A secondary goal is to provide the reader of RPO with an automatically generated scansion of a poem upon which the reader can build or criticize. The art of basic scansion is dying: more and more undergraduate students, and even graduate students of English literature, do not know where to begin when faced with the task of scanning a poem's rhythms. If RPO can present its readers with a computer-generated scansion of any poem, then perhaps the reader of the site will learn more about scansion by identifying those places in the scansion where he or she disagrees with the computer scansion. Any naive errors the computer commits will hopefully be apparent to someone with a limited ability for the understanding of poetic scansion. Any disagreement will hopefully result in the user's increasing insight into scansion, if only as a result of feeling superior to the computer.

If the goal were to produce a perfect scansion of a poem, then the computer could do part of the work and a human could verify the work and fix any errors or weaknesses. The results could be permanently saved, either in the RPO database or encoded according to TEI guidelines. But a scansion of a poem is often not unique: variations will arise based on who is scanning the poem or what mood he or she is in at the time. While TEI guidelines allow for the encoding of variant readings of the metre, encoding a basic reading even with variations

implies a permanent or absolute reading. AnalysePoems is built on the prosodic philosophy that a full scansion of a poem is an impermanent performance, much like the recitation of a speech or the execution of a piano sonata: there are basic interpretations upon which most will agree, but individual variations will inevitably arise. As a result, AnalysePoems produces an imperfect scansion and invites the student of poetry to find the weaknesses and alternate readings, allowing the student to learn more about scansion in the process of doing so. In spite of this philosophy, AnalysePoems is remarkably adept at scanning poems.

2 The RPO Database

The RPO database, at the time of writing, is a Microsoft Access database. The decision to place the poems in a relational database was a departmental decision: the Information Technology Services Department of the University of Toronto Libraries relies heavily on relational databases for data storage rather than XML or another system, and RPO needed to conform to its standards. Once in the database, of course, it is an easy matter to export the data in XML format as needed. The RPO database currently holds 3,162 poems with a total of 222,347 poetic lines. There are editorial notes to 14,548 individual poem lines. There are 500 poets in the database, many with extensive biographical information such as education, illnesses, and family relationships. Such information is stored in tables to allow for interesting queries such as: a list of all poets who died in the nineteenth century from causes unknown; a list of poems written by female poets with a university education; or a list of all poets with a spouse or sibling who was also a poet. Currently, the public RPO site does not allow for such queries, but the structure is in place for them, and plans indicate that the site will soon permit them.

The new RPO site was designed to allow the editor to input new poems quickly and easily. The most important decision, as far as the poetic data is concerned, was determining the basic poetic unit: the word, the line, the stanza, the whole poem, or some other division. A poem can be saved in a database as one field of a record: in this case, the details of the display of a poem are preserved as spacing characters and carriage returns, making the task of displaying the poem in a web page relatively easy. RPO, instead, breaks each poem down into poetic lines and saves the poem as a series of records of lines. As a result, information about stanza structure, line indentation, line numbering, and anything else of importance to the display of the poem has to be processed before the poem is saved in the database. Once in the database, a poem and its associated content and structure information can be retrieved, reconstructed, and served to a web browser for display.

The basic poetic unit in the RPO database is thus the line. Lines are grouped into stanzas and any user-defined levels that are required, such as cantos and books. A series of codes are used to identify new divisions in the poem as well as how the poem will be displayed. Many of these codes are automatically controlled by the computer, but others can be specified by the editor through the use of a minimal set of markup tags or a web interface screen of options (Fig. 1). For example, if the RPO editor wants a poem to be displayed with every stanza numbered sequentially, then he can use the tag <stanza display="##"> at the beginning of the first stanza before the poem is processed, or he can click the appropriate option of a web screen after the initial processing. A special code, not the actual stanza numbers, is saved in the database separate from the poem text. When the poem is retrieved, the code tells the computer how to number and display the numbering of the stanzas. Other similar codes exist for poem structure, editorial notes and glosses, and line numbering options.

3 The Lexicon Database

AnalysePoems relies on a Lexicon database, supplemental to the RPO database. It is also a Microsoft Access database and currently consists of three tables: WordList, WordListAlternate, and Rhyme. Each record in WordList includes the number of syllables in the word, the syllabic division of the word, and an identification of the heavy and weak accents. A Boolean flag is used to indicate if the

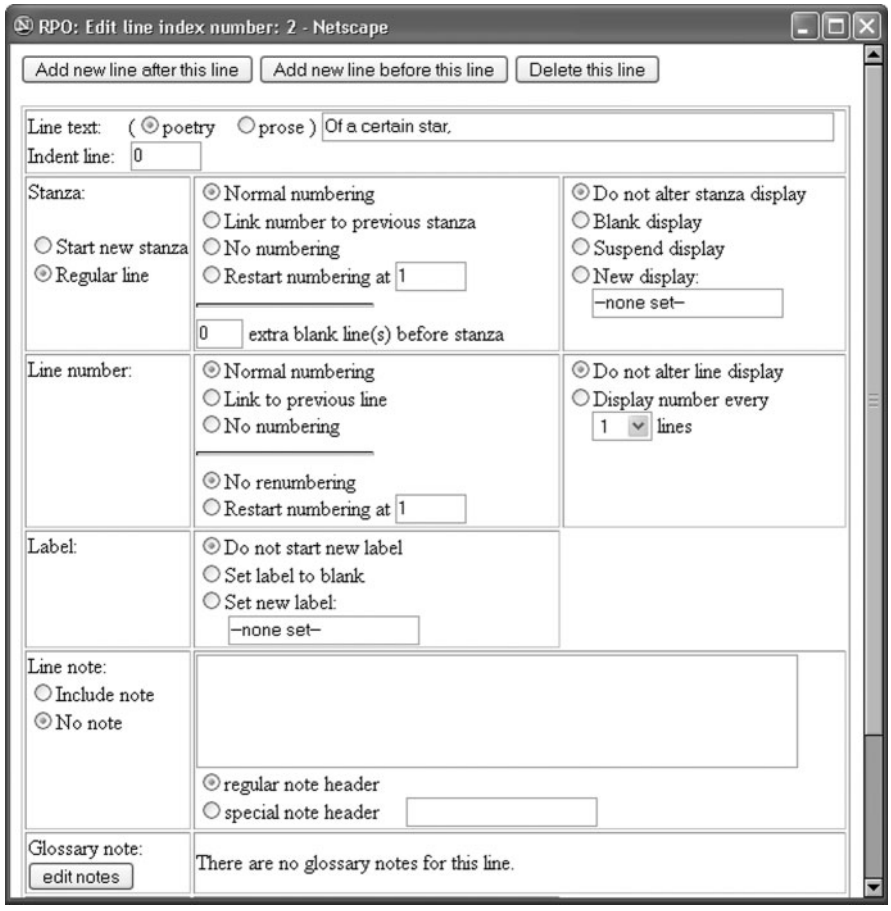

Fig. 1 A web interface option screen for editing poems in RPO

word has been verified by a human or if the data associated with the word record has been automatically generated (and thus, potentially containing errors).

Currently, words are saved semi-automatically in the Lexicon: the computer scans a poem, determines which words do not already exist in the Lexicon, and asks for confirmation to save each word. A word is defined as a unique string of characters: thus *poem*, *poems*, *poem's*, and *poems'* are four different words—the last three can be related to the first using a field devoted to establishing word equivalencies. For the sake of simplicity, hyphenated words are considered as one word: *apple*, *tree*, and *apple-tree* are three different words, *seashore* and *sea-shore* are two different words, and *monstr'-inform'-ingens-horrendous* (from Robert Browning's 'Waring') is one word. Syllabic divisions and syllabic accents must be entered manually (Fig. 2). Recording both the syllabic division and the number of syllables for each word is redundant, but useful to trap errors in the data.

The heavy and weak accents for each word are specified as integer values. Heavy accents represent strongly stressed syllables; weak accents represent weakly stressed, as opposed to non-stressed (or unaccented) syllables. Accent values are derived from a combination of dictionary guidelines for the pronunciation of words and personal knowledge of which words and syllables of words are more likely to be stressed in English poetry. For example, the word *immediately* is divided as

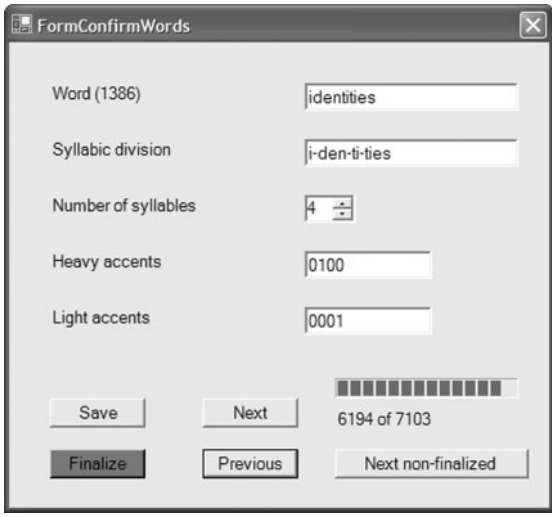

Fig. 2 Graphical interface for Lexicon word data

i-mme-di-ate-ly (note that the syllabic divisions are not standard—usually the first division will be between the two *m*s, but the division used here corresponds more closely to a pronunciation division of syllables). The heavy accent value for this word is 2, and the weak accent value is 16. These values are best understood as the binary values written in reverse: 01000 and 00010. Thus, the second syllable of *immediately* receives a heavy accent and the fourth syllable a weak accent. Heavy and weak accent values in the database represent probabilities: a syllable with a heavy accent is very likely to be scanned as a metrical stress while a syllable with a weak accent might be scanned as a metrical stress, but might not. Syllables without a heavy or weak accent will likely be scanned as unstressed syllables. Thus, there are three levels of accent—heavy, weak, and unaccented—which ultimately are reduced to only two levels of accent—stressed and unstressed—to determine the dominant metre of the poem.

The distinction between heavy and weak accents is especially important when considering monosyllabic words. Most are identified in the Lexicon as receiving a heavy accent. Others, such as pronouns, articles, and common conjugations of certain verbs (such as *to be*) are either identified with a weak accent or with no accent. The choice was based on my own conception of the probability of the monosyllabic word being stressed or unstressed in a poem. Obviously, *a*, *an*, *the*, *by*, *of*, and similar words are rarely stressed in a poem, and thus, these words receive neither a heavy nor a weak accent. Words such as *was*, *asks*, *took*, and *I* may or may not be stressed, and thus, these words receive a weak accent. A few multisyllabic words have been identified with weak accents only: *half-past* has no heavy accent and two weak accents. This allows the computer to place the stress on one or the other syllable based on the context of the poetic line.

The WordList allows for some words to have no syllabic weight. The word *i'* (a contracted form of *in*) is given a syllables value of 0. Thus, the phrase *lie i' the shad-* from D. G. Rossetti's 'The Blessed Damozel' ('We two will lie i' the shadow of / That living mystic tree') has only three syllables, not four. Similarly, *th'* has no syllabic weight. However, when the two occur together, as when Robert Herrick tells us to 'Trust to good verses then; / They only will aspire, / When pyramids, as men, / Are lost i' th' funeral fire' (in 'To Live Merrily, and to Trust to Good Verses'), they together make up an unstressed syllable.[3]

The WordListAlternate table is similar to the WordList table. It contains entries for alternate pronunciations of the words found in the main table. For now, only alternate syllabic divisions and syllabic stresses are saved in the alternate table, though eventually it will contain alternate phonetic pronunciations. The word *ignorant* has three syllables in the main table and only two syllables (*ig-n'rant*) in the alternate table. The words *power* and *our* both have two syllables, but they both have only one syllable in the alternate table. The word *subject* in the main table has a heavy accent on the first syllable (corresponding to its pronunciation as a noun); the same word in the alternate table has a heavy accent on the second syllable (corresponding to its pronunciation as a verb). So far, no word has more than one alternate pronunciation, though multiple alternate pronunciations are possible and will be taken into account by the program should they occur.

4 AnalysePoems

The AnalysePoems program was developed independently of the RPO site, but uses a copy of the RPO database.[4] While it could potentially read poems from other sources, such as text files, its current reliance on the RPO database is one of convenience: the poems in RPO, already preprocessed, save some processing time and programming effort for AnalysePoems. The program is written in Visual Basic in the .NET framework. It has various components, including one that allows easy data entry or verification of word records in the Lexicon and one that displays poet and poem information. The main parts of the program are those that analyse the metre and rhyme of poems.

4.1 Metre

A treatment of English poetic metre in current scholarship usually requires a lengthy defence. Briefly stated, the most contentious approach to metrics used by AnalysePoems is a reluctance to divide a line into metric feet. Only one test uses the usual notion of metric feet to attempt to regularize the metre; otherwise, the line is never divided into feet, except optionally at the final display. A few problems present themselves to the prosodist when considering metric feet, one of which is deciding whether or not a foot should be permitted to span a caesura, including a mid-line end-stop. The program as yet does not take caesuras into consideration, but because it also mostly stays away from feet, this is not an important problem. Also contentious, the program avoids regularizing the number of stressed syllables in a poetic line, such as the usual five stresses in a line of iambic pentameter. Forcing a line to have the expected five stresses, no more and no less, can lead to some problematic readings. The program is able to identify iambic pentameter as the dominant metre of a poem even if little more than half the lines have five syllabic stresses, while the rest have four or six.

AnalysePoems attempts as much as possible to identify metric patterns rather than impose them. Hayward's approach is to push the iambic line gently towards an iambic reading by building into the line prior to computer execution a bias towards an iambic reading (1991, p. 307). This seems reasonable enough when one knows ahead of time that all the lines being analysed are iambic or variations of an iambic line. Because AnalysePoems has no foreknowledge of the metre of any given poem, it tries first to search for a dominant pattern, and then, if one is found with a reasonable degree of confidence, it attempts to push the lines gently into the pattern.

The analysis consists of a series of tests. Each test may or may not be performed, depending on the results of the previous tests. Certain tests are repeated after other tests, the repetition of which may or may not change the results. A few confidence factors are used to determine whether or not a test may be performed.

The first thing the program does in analysing the rhythm of a poem is to divide the line into words and syllables, assign the predetermined accents from the Lexicon to the syllables of each word, and then attempt to scan the poem. In doing so, it considers an accent of 0 or 1 as an unstressed syllable and an accent of 2 as a stressed syllable. For each line of the poem, it counts the number of unstressed syllables before the first stressed syllable of the line and the number of unstressed syllables after the final stressed syllable. It also counts the number of unstressed syllables that occur between two consecutive stressed syllables throughout the poem. It assembles frequency statistics, returning the most common and the second most common values for each of the three conditions. The results are then used to determine the dominant metre of the poem. Six dominant metres are possible. Table 1 shows the values of before, after, and between and the metre that results from the combinations. If the trio of values does not correspond to any of the six combinations, then the computer decides it cannot yet identify the metre. Other metres (such as cretic) are possible, but not probable and thus, are ignored.

Associated with the identification of the dominant metre is a confidence factor. This is based on the number of unstressed syllables occurring between stressed syllables (the before and after values are not used for its calculation): it is an

Table 1 Common metrical patterns and their division of stresses

	before	after	between
Iambic	1	0	1
Trochaic	0	1	1
Trochaic catalectic	0	0	1
Anapaestic	2	0	2
Dactylic	0	2	2
Amphibrachic	1	1	2

Table 2 Initial metrical results and confidence values for twelve test poems

	rhythm	between	total	confidence
Amy Margaret's Five Years Old (Allingham)	duple	23	29	79
Dover Beach (Arnold)	duple	49	101	49
My Last Duchess (Browning)	duple	66	133	56
To the Memory of Mr. Oldham (Dryden)*	duple	30	53	61
'Out, Out –' (Frost)	duple	39	94	48
To Anthea (Herrick)*	duple	27	41	71
Ode to a Nightingale (Keats)	duple	123	239	51
When I consider how my light is spent (Milton)*	duple	25	45	60
A Daughter of Eve (C. Rossetti)	duple	17	28	66
Sonnet CXXX (Shakespeare)*	duple	12	27	44
Ulysses (Tennyson)*	duple	106	182	63
Out of the Cradle Endlessly Rocking (Whitman)*	duple	209	589	35

evaluation of the probability that the poem has either duple or triple metre (the first three metres in Table 1 are duple; the second three are triple). The confidence factor is the ratio of the number of occurrences of the most common number of unstressed syllables between two stressed syllables to the total possible number of occurrences.[5] Table 2 shows the computer-calculated confidence factor and the values used to calculate the factor for twelve poems that the computer identifies as having duple rhythm.

An asterisk in Tables 2, 3 and 5 indicates that the poem contains words that are not yet in the Lexicon. The computer is thus dealing with unknown factors and trying to find the dominant patterns in spite of the missing data (see Fig. 3 for a list of the unidentified words in Shakespeare's sonnet). Note also that the value of 56 for 'My Last Duchess' does not correspond to the ratio of 66/133, because the confidence factor includes an adjustment for the case when the second greatest number of occurrences of unstressed syllables between two stressed syllables differs from the number corresponding to the greatest number of such occurrences by two for duple metre or three for triple metre. For 'My Last Duchess,' there are sixty-six occurrences of one unstressed syllable between two stressed syllables and thirty-six occurrences of three unstressed syllables between two stressed syllables. Since, the first value represents the highest number and the second value represents the second highest number, the computer uses a quarter of the thirty-six occurrences when calculating the confidence factor: thus, $(66 + 36/4)/133 = 56\%$. This calculation is based on the assumption that at least a quarter of the time that three unstressed syllables occur between two stressed syllables, the second of the three will be promoted to a stressed syllable. This is a conservative though useful assumption to make. A similar calculation is used for the above poems by Dryden, Frost, Herrick, Milton, Rossetti, and Tennyson.

The program uses the arbitrary value of 70% for a confidence threshold: if a confidence value is less than 70%, then the computer is reluctant to identify the metre. Only two of the above twelve poems can be identified as duple metre before any further analysis, though Herrick's poem is on the threshold. The very low result for Whitman's poem should immediately signal that the poem is not written in duple metre (as indeed it is not), and since the computer identifies duple metre as the most probable metre for the poem, the computer can conclude that the metre of the poem is irregular, though it will continue to attempt to analyse the metre of the poem before deciding this.

The next step promotes syllables with an accent value of 1 (that is, a light accent) that occur where a heavy accent is expected, based on whether the metre is duple or triple. The position of the syllable

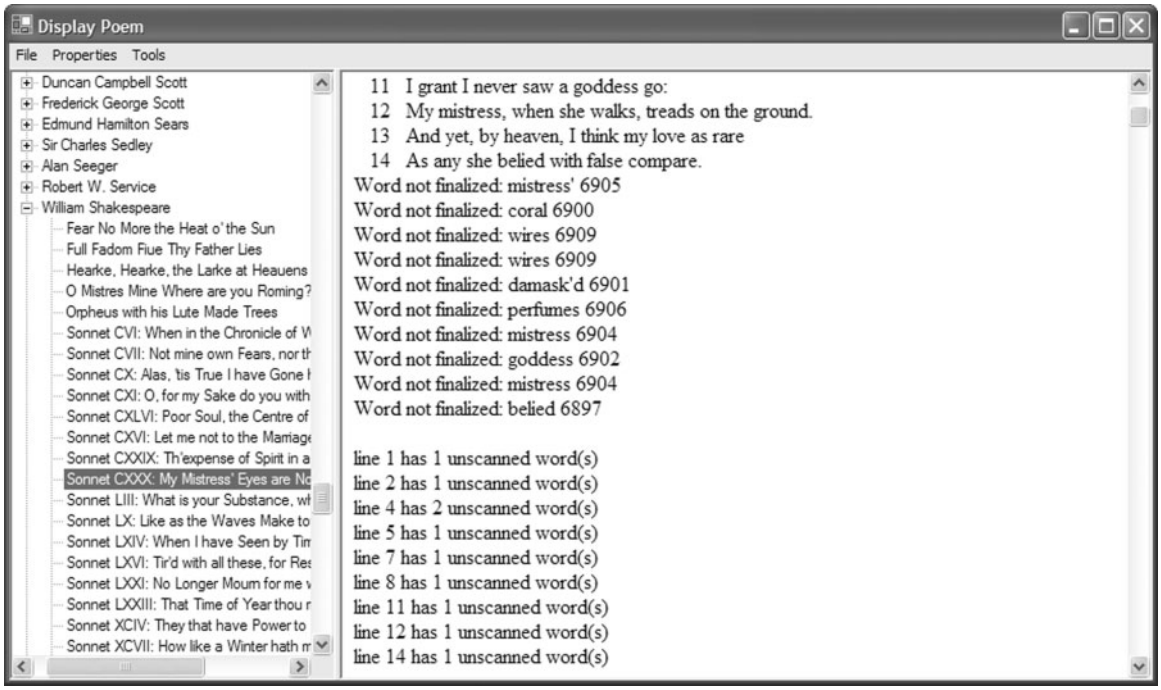

Fig. 3 Unknown words in Shakespeare's Sonnet CXXX

within the line is not used for this calculation; instead, the computer counts the number of weak syllables (accent values of 1 or 0) between two heavy accents, wherever they occur in the line. This procedure is conservative in that it will not promote syllables that have not previously been identified with a light accent, and thus, words such as articles and common conjunctions will not be promoted at this stage.

The computer then refreshes the confidence value and proceeds to examine all three- and four-syllable words in the poem whose last two syllables are unaccented and have one heavy accent (corresponding to 100, 1000, or 0100). It then promotes an unaccented syllable of the word to an accent value of 2 if the syllable occurs where such a stress is expected (resulting in 101, 1010, or 0101 for duple metre). It will do this only if the rhythm confidence is greater than 60%. Thus, the syllable *-et* in *Margaret* (when used as a tri-syllabic word) will be promoted if the conditions are met.

Table 3 displays the confidence factors for the same twelve poems after these two initial tests,

Table 3 Confidence values for test poems after initial metrical promotions

	between	total	confidence	difference
Amy Margaret's Five Years Old (Allingham)	32	37	86	+7
Dover Beach (Arnold)	75	116	65	+16
My Last Duchess (Browning)	133	173	77	+21
To the Memory of Mr. Oldham (Dryden)*	54	69	78	+17
'Out, Out –' (Frost)	87	123	71	+23
To Anthea (Herrick)*	45	52	87	+16
Ode to a Nightingale (Keats)	197	288	68	+17
When I consider how my light is spent (Milton)*	40	54	74	+14
A Daughter of Eve (C. Rossetti)	33	38	87	+21
Sonnet CXXX (Shakespeare)*	22	34	65	+21
Ulysses (Tennyson)*	172	225	76	+13
Out of the Cradle Endlessly Rocking (Whitman)*	289	643	45	+10

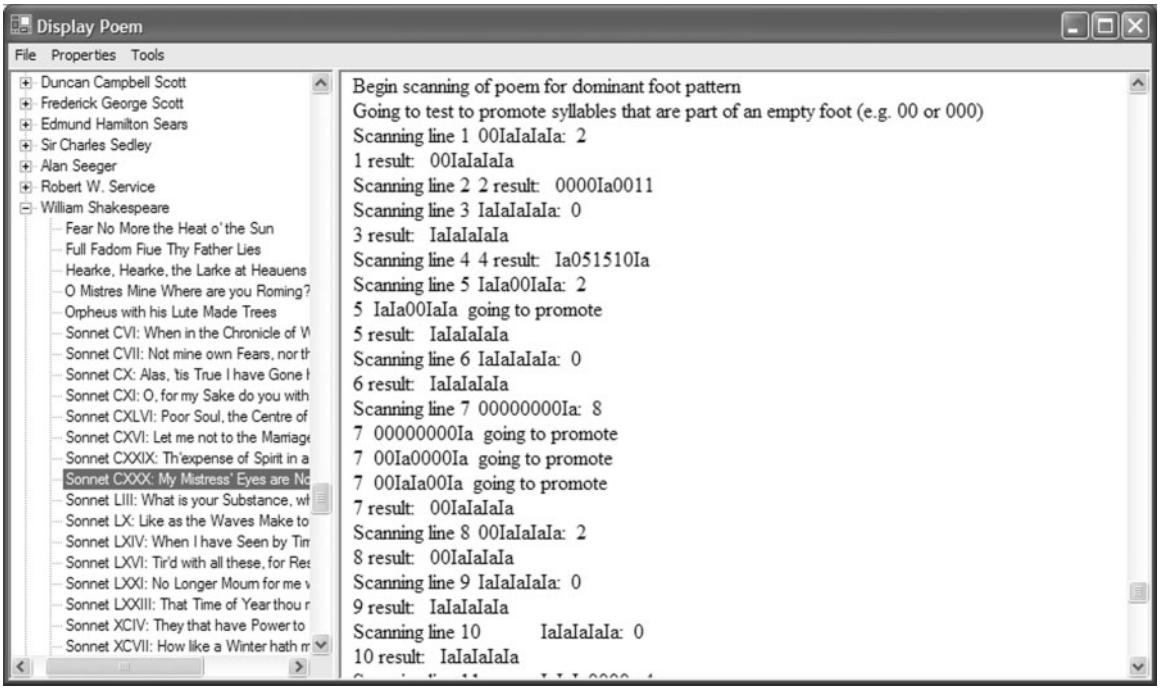

Fig. 4 Syllable promotion based on metrical foot pattern in Shakespeare's Sonnet CXXX

including the difference between these results and the results prior to these first two attempts at regularizing the metre. Note that the smallest increases in confidence correspond to the poems with the originally highest and lowest confidence values. A great increase resulting in a high confidence value corresponds to poems that are metrically regular but may not appear so at first. Such poems, with the exception of Rossetti's, are less obviously regular at first reading: they are more removed from the 'sing-song' style of rhythm found in the Allingham and Herrick poems.

After these tests, four main tests are executed. These tests are each performed more than once: they can be performed multiple times, though I have found that twice each is usually sufficient. The first test is the only time the computer attempts to scan for regular feet (Fig. 4). It searches a line for the number of occurrences of the dominant foot pattern, if it has been identified. If it finds, say, two iambs followed by two unaccented syllables followed by two more iambs and if the confidence level is high enough, it promotes the second of the two unaccented syllables thereby creating a perfectly regular iambic pentameter line. If iambic metre is the dominant metre, the computer will not promote two unaccented syllables found at the beginning of a line: another test will try to determine whether these syllables should be converted to an iamb or a trochee. This is because trochaic substitutions are fairly common as the first foot of an iambic, especially pentameter, line.

The computer then performs another attempt at pattern recognition: it tries to identify line groups based on the number of syllables in a line. This is where the computer begins to achieve a greater understanding of the structure of the poem. Consider, for instance, Herrick's 'To Anthea, Who May Command Him Anything': it consists of six stanzas of four lines, where eight-syllable lines alternate with six-syllable lines. The computer iterates through various combinations of the lines, counting the number of syllables in each line according to each test grouping, and then determines a confidence factor representing whether

a particular grouping is a recurring pattern in the poem. For Herrick's poem, the computer determines, based on the confidence levels, that the basic line grouping consists of two lines: the first of which has eight syllables and the second of which has six. There are confidence factors associated with these results: for Herrick's poem, the eight-syllable lines have a confidence of 100%, the six-syllable lines have a confidence of 90%, and the grouping of two lines together has a confidence level of 95%. For Browning's 'My Last Duchess,' the computer recognizes a line grouping of one: each line, it calculates, has ten syllables, and it is confident of this at a factor of 94%. For Whitman's 'Out of the Cradle Endlessly Rocking,' it identifies a line grouping of three (of six, thirteen, and twelve syllables), but is confident of this at a factor of 12%: in other words, it finds no pattern. For Keats's 'Ode to a Nightingale,' the computer recognizes that the lines occur in groups of ten.[6] Table 4 presents the results of the computer's analysis of Keats's poem: the first set results from a test before alternate pronunciations are tested, and the second set from after the test. One of the interesting results is the tenth line of the group: half of the final lines of Keats's eight stanzas have eleven syllables (the other half is made up of three occurrences of ten syllables and one occurrence of twelve syllables). Keats took more licence with the final line of his stanza than with any of the other nine lines in the stanza.

It is important for the computer to recognize the number of syllables in each line so that it can proceed to the next stages: measuring the syllabic length of unknown words and testing for alternate pronunciations. If a high enough confidence factor is achieved from the line groupings, then the computer will be able to deduce the syllabic length of unknown words, so long as only one word occurs in each line under consideration (or the same word occurs twice). This stage is remarkably accurate. In Milton's sonnet 'When I consider how my light is spent,' two words were deliberately left out of the Lexicon: *therewith* and *day-labour*. Because the program is 100% confident that all the lines of Milton's poem have ten syllables (note that lines with unknown words are not taken into consideration when tabulating line grouping results), it deduces that *therewith* has two syllables and *day-labour* has three. Whitman's poem has four unknown words, but the computer can do nothing with them because the confidence factor for line grouping is so low. In Dryden's 'To the Memory of Mr. Oldham,' eighteen words were left out of the Lexicon. The computer correctly identifies all of them except *write* and *encompass*, which both occur in lines of twelve syllables instead of the regular ten syllables. The word *generous* is interesting: the computer identifies this word as having two syllables instead of the usual three, but *gen'rous* is an acceptable alternate pronunciation of the word, and was probably the pronunciation Dryden was using for this poem.

Table 4 Confidence values for identifying the number of syllables in each line of a grouping before and after alternate pronunciations are tested in Keats's 'Ode to a Nightingale'

	syllables	**confidence**	**syllables**	**confidence**
line 1	10	75	10	75
line 2	10	75	10	87
line 3	10	62	10	100
line 4	10	87	10	100
line 5	10	100	10	100
line 6	10	75	10	87
line 7	10	50	10	75
line 8	6	75	6	100
line 9	10	75	10	100
line 10	11	50	11	50
total		72		87

The final main test is the test for alternate pronunciations. All words in the poem are replaced in turn by each of their alternate pronunciations. A few tests evaluate whether to accept or reject an alternate pronunciation for the word in question: syllable length of the line and regularity of metre are the two main tests. The alternate pronunciation test is performed only if the line grouping confidence and the rhythm confidence are sufficiently high (greater than 70%). This is a successful routine: the program determines when to replace *power* by *pow'r* (of course, *pow'r* will never be replaced by *power*) and *minute* (with an accent on the

first syllable) by *minute* (with an accent on the second syllable). Something that it does not test, but probably should, is alternate pronunciations in varying combinations. At the moment, one alternate pronunciation is tested at a time: if, say, there are three words in a line each with one alternate pronunciation, the computer should go through all seven possible alternate combinations and choose the most successful. The tests that decide on the most successful combination, though, must be more rigorous, and are possibly more trouble than they are worth.

Aside from these four main tests, other minor tests are used. The elision of the word *the* is checked: if the word occurs before a word beginning with a vowel, then *the* may possibly be elided with the following word. Also, the beginnings of predominantly iambic lines are checked for trochaic substitutions (if the line begins with two syllables of which neither receives an accent value of 2). Both these tests are reasonably successful, though the second especially produces an occasional misreading.

The computer periodically attempts to scan the metre of the poem, updating the confidence factor. The tests are performed more than once: sometimes a test must be skipped the first time because the confidence factor is too low. A different test may raise the factor and the test can be performed at the next attempt. In this manner, the scansion of the poem is gently nudged towards what the computer identifies as the dominant pattern. If the metre is regular enough, the computer can then scan unknown words. For the two unknown words in Milton's sonnet, the computer identifies *day-labour* as having a heavy accent on the second syllable (*la-*) and *therewith* as having a heavy accent also on the second syllable (*-with*). Now that the computer knows the syllabic length of the words and where the heavy accents are, the words are saved in the Lexicon. The words are saved with the finalized flag set to false so that a human can go back and approve or edit these and all words that the computer entered automatically.

An interesting example poem is Shakespeare's Sonnet CXXX ('My mistress' eyes are nothing like the sun'). Eight of its words were deliberately left out of the Lexicon: *mistress'*, *coral*, *wires*, *damask'd*, *perfumes*, *mistress*, *goddess*, and *belied* (Fig. 3). After the initial tests, the rhythm confidence jumps from 44% to 65%. After going through three of the four main tests once, the program is able to identify the syllabic length of *coral*, *mistress*, and *belied*: all with two syllables. The rhythm confidence then jumps to 81%. It then decides to use the alternate pronunciation of *heav'n* and goes through the main tests again. This time it identifies the syllabic lengths of the other words: *wires* as monosyllabic and the rest as disyllabic. (The monosyllabic pronunciation of *wires* will eventually be placed in the alternate pronunciations list, giving priority to the disyllabic pronunciation.) The computer's final analysis is that the poem is written in iambic metre (with an 83% confidence factor for the rhythm) and that every line has ten syllables. The final scansion of the poem is the following:

my(1) MISTRESS'(2) –(0) eyes(2) are(1) no(2)
thing(0) LIKE(2) the(0) sun(2)
coral(0) –(0) is(1) far(1) more(1) red(2) than(0)
her(0) lips'(2) red(2)
if(0) snow(2) be(0) white(2) why(1) THEN(2)
her(0) breasts(2) are(1) dun(2)
if(0) hairs(2) be(0) WIRES(2) black(2)
WIRES(2) grow(2) on(0) her(0) head(2)
i(1) HAVE(2) seen(1) ro(2) ses(0)
DAMASK'D(2) –(0) red(2) and(1) white(2)
but(1) NO(2) such(1) ro(2) ses(0) see(2) i(1)
IN(2) her(0) cheeks(2)
AND(2) in(0) some(1) PERFUMES(2) –(0) IS(2)
there(1) MORE(2) de(0) light(2)
than(0) in(0) the(0) breath(2) that(1) FROM(2)
my(1) MISTRESS(2) –(0) reeks(2)
i(1) love(2) to(0) hear(2) her(0) speak(2) yet(1)
WELL(2) i(1) know(2)
that(1) mu(2) sic(0) HATH(2) a(0) FAR(2)
more(1) plea(2) sing(0) sound(2)
i(1) grant(2) i(1) ne(2) ver(0) SAW(2) a(0)
GODDESS(2) –(0) GO(2)
my(1) MISTRESS(2) –(0) WHEN(2) she(1)
walks(2) treads(2) on(0) the(0) ground(2)
and(1) yet(1) by(0) heav'n(2) i(1) think(2)
my(1) love(2) as(1) RARE(2)
as(1) a(2) ny(0) SHE(2) belied(0) –(2) with(0)
false(2) com(0) pare(2)

A word or syllable in capital letters indicates a promoted syllable (one that was originally unaccented, with a light accent, or unknown). A dash indicates the presence of a syllable when the computer is unsure how the word divides. The numbers in parentheses are the final stress values of the words: 0 and 1 correspond to unstressed syllables and 2 corresponds to stressed syllables. The openings of lines are less regular than the closings of lines, partly as a result of a reluctance to impose the regular rhythm on the first few syllables. Consider 'And in some perfumes' and 'Than in the breath.' The computer promotes the *and* because the word is identified in the Lexicon as having a weak accent: it thus carries a greater probability of promotion than *in*. Whether *and* or *in* should be promoted is contingent on the performer/reader of the poem. In the following line, *in* should probably be promoted—and it would be if the phrase were to occur anywhere in the line but the very beginning—but, the reluctance to promote at the beginning of a line results in no promotion. A performer of the poem may decide to emphasize *than*, *in*, or neither—and thus, the opening of this line is best left alone.

The scansion of the poem is not perfect, but there are no glaring errors. The second and the fourth lines are the most irregular: the computer especially does not know what to do with the second line because the opening word is originally unknown and the final four syllables are not iambic. But what Hartman claims of his 'Scansion Machine' (1996, p. 42)—that the computer scans poems at about the same level as a good undergraduate university student after a few months' experience in scanning poems—can be used to describe AnalysePoems. In fact, a student may be tempted to scan 'My mistress, when she walks, treads on the ground' with a stress on *on* and no stress on *treads* due to the iambic pull of the line: my preference is to stress *treads* and not *on* which is also what the program decides to do. AnalysePoems knows to avoid the more naive scansion.

Table 5 shows the final confidence factors for the same twelve poems, along with the difference between the final factor and that after the first two initial tests and the difference between the final factor and that of the initial scansion, before any tests. These results suggest that a value for the confidence factor in the 80s denotes a regular metre, a value in the 70s denotes a regular metre with many variations, and a value less than 50 or 60 denotes an irregular metre. The biggest change in the confidence factor, for Shakespeare's sonnet, is partly a result of the greater proportion of unknown words in the poem compared to the other poems. The poems with the most obvious 'sing-song' quality were the least affected by the four main tests, though Frost's and Whitman's poems were very little affected and neither can be characterized as having a 'sing-song' rhythm.

Table 5 Final metrical confidence factors for the test poems with confidence value changes

	confidence	difference	difference
Amy Margaret's Five Years Old (Allingham)	86	0	+7
Dover Beach (Arnold)	69	+4	+20
My Last Duchess (Browning)	83	+6	+27
To the Memory of Mr. Oldham (Dryden)*	84	+6	+23
'Out, Out –' (Frost)	73	+2	+25
To Anthea (Herrick)*	87	0	+16
Ode to a Nightingale (Keats)	73	+5	+22
When I consider how my light is spent (Milton)*	79	+5	+19
A Daughter of Eve (C. Rossetti)	87	0	+21
Sonnet CXXX (Shakespeare)*	83	+18	+39
Ulysses (Tennyson)*	84	+8	+21
Out of the Cradle Endlessly Rocking (Whitman)*	47	+2	+12

4.2 Rhyme

The rhyme analysis of AnalysePoems is much simpler than the metrical analysis. A poem, though, must be scanned for metre before its rhyme can be analysed. This is because full rhymes occur on emphasized syllables. Rhymes on unstressed syllables are half rhymes. In Browning's

'Youth and Art,' the words *beard too* rhyme with *adhered to* not because *too* rhymes with *to* but because *beard too* rhymes with *-hered to*. The words *subject* and *reject* rhyme if they are both verbs, but do not rhyme if one or both are nouns. The emphasized syllable helps determine which words rhyme with which.

Rhymes are saved in the Lexicon as rhyme groups. When the computer identifies a new rhyming pair, it checks to see if either word already exists in the rhyme table: if so, and the other word does not, it assigns the rhyme group number of the first word to the second word and saves the data. If neither word is in the table, it assigns a new rhyme group number to the pair and saves the data for both words. This way, rhyme groups gradually increase in population. At times, human input is needed to join two rhyme groups that the computer has not yet realized should form but one group.

The computer begins building the Rhyme table by examining poems whose rhyme scheme has already been identified by RPO. Once it has found a sufficient number of rhyme words, it can then analyse the rhyme pattern of poems with no specified rhyme scheme. Knowing the syllabic line grouping and the number of lines to a stanza helps the computer test rhyme schemes. It determines the rhyme pattern in much the same way it determines line grouping: it tests a few options and decides which is most probable. It has some trouble distinguishing *abcb* quatrains from *abab* quatrains, but otherwise it works quite well. Human input is needed to correct and finalize some of the extrapolated data, but as the list of rhyme words increases in size, corrections become less and less frequent.

5 Conclusion and Future Work

AnalysePoems is very successful at identifying the dominant metre of a poem written in accentual and accentual-syllabic metre, the dominant metres of English poetry from at least the sixteenth through at least the nineteenth centuries. It is able to do so knowing only what most readers of poetry know when first approaching a poem: the way a word divides into syllables and the usual placement of accents on those syllables. It is even able to identify the metre without knowing such things for all the words in the poem, and it can then usually deduce these for the unknown words. It can produce a scansion of the poem: one that is not perfect, but one that should be useful to readers of poetry who have difficulty knowing how to begin scanning the metre of a poem. It has also begun learning to identify the rhyme schemes of poems, easily distinguishing between rhyming and non-rhyming poems.

The computer has not yet attempted to scan all the poems in the RPO database: a few more refinements are needed first. Yet to be implemented is the ability to distinguish between accentual-syllabic and syllabic verse and other verse types such as Hopkins's sprung rhythm. Also desirable is the ability to identify when the lines of a poem are not written with the same metre. For example, Houseman's 'On the idle hill of summer' has alternating eight- and seven-syllable lines: the first line of the two is trochaic while the second is trochaic catalectic (the final unstressed syllable is omitted). More difficult is the ability to identify less traditional metres, such as the interplay of iambs and anapaests in Browning's 'Prospice'—as it is, the computer identifies the metre as anapaestic with a confidence factor of 57%, due to the fact that while about half of the feet are anapaests, the other half are iambs.

The phonetic pronunciation of words needs to be incorporated into the Lexicon. This should enhance the computer's ability to identify rhythmic effects, such as slight pauses between words that are not signalled by punctuation. It will allow the computer to identify alliteration, which has an impact upon rhythmic stress. It will also allow the computer to identify alliterative stress verse, common in Anglo-Saxon poetry. It might also reveal less obvious phonetic patterns in poems, such as evenly-spaced, recurring vowel sounds. A phonetic scansion might reveal hitherto unnoticed aspects of poetic aesthetics. A computer analysis of poetic aesthetics, both metrical and phonetic, is the ultimate goal of AnalysePoems.

AnalysePoems demonstrates that the automation of a basic scansion of regular rhythm poems

is possible. Most poems do not have a perfect or absolute scansion: this is perhaps too often ignored. Programmers should not be concerned with getting a computer to scan a poem flawlessly: interpretational variations of a poem's metre must be allowed. This is why AnalysePoems is more concerned with identifying the dominant metrical pattern rather than producing a perfect scansion. The scansion that does result, though, is usually accurate in the right places and interpretationally reticent in the right places. Confidence values are useful for overcoming the problems of these interpretational variations as well as incomplete lexical data. The use of confidence values helps identifying patterns without imposing preconceived patterns when there is no justification for doing so. Initial results suggest that these confidence values actually yield insight into the rhythmic makeup of a poem: not only do they help identify the dominant rhythm, they point to the relative simplicity or complexity of the metre. Allingham's 'Amy Margaret's Five Years Old' and Browning's 'My Last Duchess' have both been identified with duple metre, the one with 86% confidence and the other with 83% confidence. The fact that Browning's poem's initial confidence value was 27% lower than the final value while that of Allingham's poem was only 7% lower suggests that 'My Last Duchess' has more complex poetic rhythms than 'Amy Margaret's Five Years Old.' This greater complexity results, some might argue, in a more rewarding reading experience. Thus, the value of 27% is a numerical insight into the aesthetics of Browning's poem.

References

Hartman, C. O. (1996). *Virtual Muse: Experiments in Computer Poetry*. Hanover, New Hampshire: Wesleyan University Press.

Hayward, M. (1991). A connectionist model of poetic meter. *Poetics*, **20**: 303–17.

Hayward, M. (1996a). Analysis of a corpus of poetry by a connectionist model of poetic meter. *Poetics*, **24**: 1–11.

Hayward, M. (1996b). Applications of a connectionist model of poetic meter to problems in generative metrics. *Research in Humanities Computing: Selected Papers from the ALLC/ACH Conference*, **4**: 185–92.

Lancashire, I. (2002). Free poetry for the world: editing representative poetry online. *Journal of Scholarly Publishing*, **34**(1): 16–29.

Logan, H. M. (1982). The Computer and metrical scansion. *ALLC Journal*, **3**(1): 9–14.

Representative Poetry Online. Ian Lancashire (ed.). Available from: http://rpo.library.utoronto.ca/display/index.cfm (Accessed 28.12.05).

Notes

1 The new RPO site (version 3) was launched on 16 October 2002. While I was the main programmer for the site, the design, both visual and structural, is a result of a collaboration with Ian Lancashire and Sian Meikle. Alan Darnell was especially helpful with design suggestions in the early stages of development.
The web site is programmed in ColdFusion. When first launched, the new website produced HTML pages on demand: no static HTML files were saved for the poets, poems, and indices. As the site grew in popularity, the number of requests for poems and other dynamically-generated pages overwhelmed the server. The site was then quickly converted to a collection of ColdFusion-generated static HTML pages: the HTML pages are automatically updated as required, such as when a poem is edited or newly added.

2 For these reasons and others, I largely use traditional metric scansion, instead of generative metrics.

3 All poetic quotations are taken from *Representative Poetry Online* (http://rpo.library.utoronto.ca).

4 At the moment, there are no plans to implement AnalysePoems on the web site, though the program was developed with the RPO web site in mind.

5 The confidence factor can be interpreted as a measure of the rhythm's regularity, since it is a percentage of occurrence of either duple or triple rhythm. If I were to call it a regularity factor, I might be suggesting that regular rhythm is an important aspect of a poem, when indeed variations in the rhythm are not only permitted in most poems, but are often interesting aspects of the poem. It must be noted, though, that a confidence factor of 100% is neither common nor necessary when evaluating the dominant rhythm.

6 Note that the RPO database saves information about the stanza structure of poems, though only implicitly. It records a stanza break when it encounters a blank line.

It saves this information as a special code, but does not attempt to count the number of lines in each stanza, nor will it signal an irregularity as a possible error. AnalysePoems can easily determine the stanza length for each poem based on this information, but here it is concerned with line groups instead. With Herrick's poem, the stanza length is four lines but the smallest line grouping is two lines; with Keats's poem, the smallest line grouping and the stanza length are the same.