Ranking Contemporary American Poems

Michael Coleman Dalvean

Australian National University

School of Politics and International Relations

michael.dalvean@anu.edu.au

January 2013

Introduction

The purpose of this paper is to examine what distinguishes a "professional" poem from an "amateur" poem. The central idea here is that professional poets are more likely than amateur poets to have grasped the basic skills associated with writing poetry and have therefore been able to produce poems of lasting quality. Amateurs, on the other hand, are less likely to have mastered the basic required skills and are therefore less likely to have produced work of lasting quality. Intuitively, we know that there are differences between the skills of amateurs and professionals in various fields and we are quick to make aesthetic judgments based on our raw subjective responses. However, the objective quantification of the factors that lead to such responses is rarely considered. By using computational linguistics it is possible to objectively identify the characteristics of professional poems and amateur poems. This way an objective basis for our subjective responses can be identified.

The upshot of identifying the characteristics of high quality poems is that we can then come up with a means of placing poems on a continuum according to how much a poem exemplifies the characteristics of an amateur poem or, at the other extreme, a professional poem. We can then use this continuum to rank professional poems and, in doing so, we can make some objective statements about which poems are "better". There is a tradition of considering some poets as "minor" and others as "major" (Eliot, 1946). Placing poems on a continuum that is based on the extent to which poems possess the craftsmanship of a professional may be a step towards explaining why some poets are "greater" than others. Thus, an important element of this paper is the creation of such a continuum using a corpus of contemporary American poets.

Related Work in Computational Linguistics

Several computational linguistic approaches to the analysis of poetry have been made. Rhyme and meter have been quantified (Green, Bodrumlu, & Knight, 2010) and methods to classify poems according to individual authors and styles have been used (Kaplan & Blei, 2007). However, only two attempts have been made to isolate the variables associated with poetic talent. The first study to use computational linguistics to identify high quality poetry is Forsythe (Forsythe, 2000) which looked at the characteristics of English poems over the last 400 years. The analysis here was based on a study group of poems that consistently appeared in recent anthologies. A control group was selecting an "obscure" poem initially published in the same year as one of the poems in the study group. The obscure poems had not subsequently appeared in an anthology. This resulted in a sample consisting of 85 "successful" poems and 85 "unsuccessful" or "obscure" poems matched by year of publication. The study found that the successful poems had fewer syllables per word in their first lines and were more likely to have an initial line consisting of monosyllables. It was also found that successful poems had a lower number of letters per word, used more common words, and had simpler syntax. Thus, contrary to what we might expect, the more successful poems used simpler language. In essence, poems that use language that is simple and direct are more likely to be reproduced in anthologies. The second study is that of Kao and Jurafsky (2012). This study used a study group of 100 "successful" American poems, where success was defined as having been reproduced in the anthology *Contemporary American Poetry* (Poulin & Waters, 2006). They used a control group of 100 amateur poems selected from an amateur poetry website (www.amateurwriting.com). In terms of effect size and statistical significance, the biggest difference was that the professional poets used words that were more concrete than the amateur poets. Furthermore, the amateur poets were more likely to use perfect rhymes rather than approximate rhymes, more alliteration and more emotional

words, both negative and positive. Finally, professional poets tend to use a greater variety of words than amateur poets. That is, the number of different words in the 100 professional poets is greater than the number of different words in the amateur corpus. This is not to say that they use more complex words, merely that they use a greater variety of simple words.

An Alternative Approach

In this paper I attempt to extend the kind of analysis undertaken in Forsythe (2000) and Kao and Jurafsky (2012). That is, I wish to determine what distinguishes a well-crafted poem from a less well-crafted poem. I use the same data as that used by Kao and Jurafsky (2012). However I extend the analysis in two ways. Firstly, I examine a broader range of linguistic variables than Kao and Jurafsky. The significant insight from Kao and Jurafsy's (2012) analysis is that the concreteness of words is far more important an indicator of poetic quality than any of the characteristics we might usually associated with poetic craft such as perfect end rhyme frequency or the type/token ratio. Therefore, if a search is made for linguistic characteristics using the types of variables that have been investigated in relation to language processing then there is the possibility that the insights gained by Kao and Jurafsky (2012) can be further extended. For this purpose I use 68 linguistic variables derived from Linguistic Inquiry and Word Count (Pennebaker, Francis, & Booth, 2001) and 32 psycholinguistic variables from the Paivio, Yuille and Madigan (1968) word norms. It will become apparent that this approach provides a further insight into the types of linguistic characteristics that distinguish professional from amateur poems.

A second way in which I extend the analysis of Kao and Jurafsky (2012) is to use machine learning to develop a classifier. The idea here is that if there are characteristics that distinguish amateur from professional poems then it should be possible to classify a given poem as being more towards the amateur end of the spectrum or more towards the

professional end. This being the case, it is also possible to *rank* individual poems according to their position on the spectrum. Thus, given Kao and Jurafsky's (2012) selection of 100 professional poems it should be possible to rank them according to where they are on the spectrum. In this sense it is possible to state that, even among professional poets, some are better than others.

Method

The Data

The data consist of the 200 poems used by Kao and Jurafsky (2012).[1] Of these 200 poems, 100 are professional poems drawn from *Contemporary American Poetry* (Poulin and Waters, 2006) and 100 are amateur poems drawn from www.amateurwriting.com. The professional poems were written in the later half 20th century by poets who have been members of the Academy of American Poets. In the 100 poem corpus there are 67 individual poets. The number of poems chosen from the anthology was in direct proportion to the number of poems the poet had in the anthology. Where a poem was over 500 words it was removed and replaced by another poem by the same poet. The final selection of 100 poems had an average of 175 words (min = 33; max = 371) (Kao & Jurafsky, 2012, p. 4).

The 100 control poems were selected from www.amateurwriting.com which is a free website on which anyone is able to post their writing. Of the 2500 available at the time of selection, 100 were randomly selected and corrected for grammar and spelling. The average length of poems was 136 words (min = 21; max = 348)) (Kao & Jurafsky, 2012, p. 4).

The Variables

[1] I would like to thank Justine Kao for supplying me with the data used in the analysis.

The dependent variable in the analysis is a binary taking the value of 1 if the poem is by a professional poet and 0 if it is not. The independent variables are linguistic variables derived from two sources – Linguistic Inquiry and Word Count (LIWC) and the Paivio Yuille and Madigan (1968) word norms and their extension by Clarke and Paivio (2004).

Sixty eight linguistic variables were derived from Linguistic Inquiry and Word Count (LIWC). This program breaks text down into linguistic categories according to a specifically designed dictionary (Pennebaker, Francis, & Booth, 2001). The categories used are based common behavioural and cognitive processes and include Negative Emotion, Affect, Leisure, Work, Family, Social Activities and Psychological Processes. The categories were derived from lists of words empirically associated with each category. Thus, the Psychological Processes category was derived from words developed from the Positive Affect Negative Affect Scale (Watson, Clarke and Tellegen, 1988, cited in Pennebaker *et al* 2007), Roget's Thesaurus, and standard English dictionaries. Thus, with sixty-eight linguistic categories LIWC captures a great deal of the linguistic content of a given text.

An additional 32 psycholinguistic variables were derived from Paivio Yuille and Madigan's (1968) word norms and the extension of these by Clarke and Paivio (2004). The Paivio Yuille and Madison (1968) and Clarke and Paivio (2004) (PYMC) word norms are derived from a sample of 925 nouns. For each word, 32 linguistic and psycholinguistic variables were derived. Some of these are structural such as the number of letters and number of syllables. Another set of variables were derived from subjects' responses to the words by getting to answer questions on a number of psycholinguistic dimensions. The variable "meaningfulness" was derived by asking subjects, for each word, how many associated words they could think of in 30 seconds while the variable "age of acquisition" (AOA) was derived by asking subjects at what age they estimate they learnt each of the 925 words. The result is that there are 32 variables for each of the 925 words that measure their structural and

psycholinguistic properties. In order to illustrate how the poems were scored on each of these 32 variables I shall use the "ease of definition" (Def) variable. This variable was derived by asking how easy is was to define each of the 925 words on a scale of 1 (very hard) to 7 (very easy). Thus, for each of the 925 words we have a Def score. Out of the 925 word sample the word that was easiest to define was "baby" (score = 6.79) and the word that was the hardest to define was "gadfly" (score = 1.92). The average score for the 925 words was 5.14. Words with in this range were "vessel" (5.13), "warmth" (5.13), "alimony" (5.17) and "caravan" (5.17).

To use the raw Def scores to score poems, the first stage was to determine, for each poem, which of the 925 words in the PYMCP sample were present. The average Def score for each poem could then be calculated. Consider for example the sentence

"The baby ridiculed the gadfly's caravan",

In this sentence the words "the" and "ridiculed" are not in the 925 word sample so they are not part of the calculation. The remaining words, "baby", "gadfly", and "caravan", are in the sample and have scores of 6.79, 1.92, and 5.17 respectively. The sentence contains three words from the sample so the "Def" score for the sentence is calculated as follows:

(6.79 + 1.92 + 5.17)/ 3 = 4.6.

Using this methodology we get a proxy for the average Def (ease of definitions) of words used in each poem. It is only a proxy because it is based on a 925 word sample. The poems were scored on all 32 psycholinguistic variables in the same way as described above for Def.

Thus, the data consist of a corpus of 200 poems with the 100 professional poems scored as 1 and the amateur poems scored as 0. For each of these poems there are 68 linguistic variables derived from LIWC and 32 derived from the PYMCP norms.

Machine Learning

It is apparent that the number of variables under consideration is half the sample size. In traditional hypothesis testing this would be a problem. However, recent advances in machine learning have pointed the way towards making sense of situations in which there is a great number of independent variables. Much of this approach has been developed in the context of gene sequencing in which it is not unusual to have a sample size of less than 200 and yet the number of independent variables that need to be considered is several thousand. Ultsch and Kämpf (2004) give an example of a data set consisting of 72 leukemia patients and 7192 variables. Clearly there needs to be some way of selecting the variables that are likely to provide the best signal. The solution used in this paper is to use logistic regression with forward stepwise selection. Under this procedure variables are selected according to an algorithm that surveys all the independent variables and selects the independent variable that provides the best logistic fit for the dependent variable. This procedure continues until no additional variables can be found that add to the model's ability to fit the data. Clearly, this can lead to problems because it is possible that variables are selected due to their ability to learn the "noise" in the dataset rather than generalize. This is known as "overfitting" (Hawkins, 2004). To prevent overfitting, an independent holdout sample can be used to check the generalization ability of the model at each of the steps in the stepwise procedure. The idea here is that the testing sample will be "held out" from the model building procedure and will only be used to test the generalization ability of the model at each stage of its development. Typically, the generalization ability of a model rises with the first few independent variables added and then falls away as more independent variables are added. As independent variables are added the *internal* measures of model fit such as R^2 tend to rise consistently but the *external* generalization ability (that is, the ability to classify cases that were not used in the creation of the model – the "held out" cases) falls considerably after the

first few variables are selected. The idea is to choose the model that maximizes the external generalization ability.

It is important to specify the holdout sample correctly as it must at all times be separate from the sample of the data used to create the model. The idea here is that a certain proportion of the data *p* should be used to create the model and the remaining proportion 1 - *p* should be used to test that the model has not been overfitted. If the model is able to generalize then it should be able to correctly classify cases that were not used in creating it. This "holdout" sample is one way of doing this and is a standard method of testing models in machine learning.

Another technique derived from machine learning is the use of an ensemble of models to increase the classification accuracy. The idea here is that averaging the outputs of several different models will likely increase the overall accuracy. This assumes that the errors of each constituent model in the ensemble are not correlated. One way to do this is to train different models on different subsets of the data. Another way is to use different variables in each constituent model. In this paper the latter approach is the one used.

Before discussing the modeling process in detail it is worthwhile to consider a question that arises in relation to the studies that have been done with this data previously: Why not simply use the logistic equation from Kao and Jurafsky's (2012) analysis? The answer is that there is a problem with overfitting in any modeling and, although it is possible that their equation is not overfitted, in the absence of an independent test using a holdout sample or some similar method, it is always possible that the equation does is overfitted to the data. In such cases the model does not truly generalize but instead "learns" the noise in the sample and is therefore not useful for actually classifying poems into professional amd amateur. This is despite the fact that certain variables may have been identified as being important in such a classification scheme. There is a distinction between traditional

hypothesis testing and machine learning. Traditional hypothesis testing is based on the idea that the identification of statistically significant variables is the essential aim as it is required to develop theoretical explanations. The problem with such an approach is that it can lead to the identification of variables that have statistical significance but little discriminant power. The central aim of machine learning, on the other hand, is classification so the discriminant power of the variables selected in crucial. The statistical significance of variables is not as important as whether they are able to increase the classification accuracy of the model.

Modeling and Results

The first stage of the modeling procedure is to divide the sample (n=200) into a training sample of n = 100 and a testing sample of n = 100. The training sample will be used to create models using the stepwise procedure while the testing sample will be "held out" from the model building procedure and used only to test each model created at each step of the stepwise procedure. Thus, 50 of the amateur poems were randomly selected from the 100 amateur poems and 50 of the professional poems were randomly selected from the 100 professional poems.

The next stage of the process was to run the stepwise procedure using all 100 linguistic variables. The stepwise procedure continued for 13 iterations and then stopped. The best classification accuracy for the holdout sample occurred at step 2. This model consisted of two variables: article (e.g.: "the", "a") and; insight (e.g.: "explain", "feel"). Both of these are LIWC variables. The sensitivity was 76%, the specificity was 72% giving an overall accuracy of 74%. This yields a Cohen's Kappa value of .48 which is highly statistically significant (Test of Ho: Kappa=0: z=4.80, p =0.0000 t.t.t.). Parameter estimates for this model are presented in Table 1.

Table 1 about here

The next model was created by removing the two variables article and insight from the pool of potential independent variables and running the stepwise procedure again. The stepwise procedure continued for 5 iterations and then stopped. The best classification accuracy for the holdout sample occurred at step 2. This model consisted of two variables: affect (e.g.: "gentle", "terrible") and; cognitive mechanisms (e.g.: "imagine", "consider"). Both of these are LIWC variables. The sensitivity was 74%, the specificity was 74% giving an overall accuracy of 74%. This yields a Cohen's Kappa value of .48 which is highly statistically significant (Test of Ho: Kappa=0: z=4.80, p =0.0000 t.t.t.). Parameter estimates for this model are presented in Table 2.

Table 2 about here

The two variables affect and cogmech were removed from the potential pool of independent variables and the stepwise procedure run again. However, subsequent models had a lower classification accuracy than Models 1 and 2. The summary accuracy and parameter estimates for models 1 and 2 are given in Table 3.

Table 3 about here

The PYMC variables were not selected by the search procedure in the creation of the first two models. In order to introduce them into the analysis a different search procedure was undertaken. All the LIWC variables were removed from the potential pool and only the PYMC variables were retained for subsequent model building. The idea here is that the

stepwise procedure is a “greedy” search algorithm which takes, at each step, the variable with the greatest model fitting power. This means that some combinations of variables can be overlooked because some variables work best when combined with other variables which may not be identifiable with individual sweeps of the data. The model building described above did not use any PYMC variables because, as individual variables, the LIWC variables performed better. By eliminating the LIWC variables there is the possibility that some combination of PYMC variables will be selected and, in combination with other PYMC variables, perform well.

Thus, the same procedure as that enumerated above was undertaken but with only the PYMC variables. That is, when the best model for a given iteration was identified, the constituent variables from that model were eliminated from the pool of potential independent variables and the procedure was run again. The resulting models from this procedure are listed in Table 4.

Table 4 about here

Clearly, all the models created using the LIWC variables (Models 1 and 2) and those using the PYMC variables (Models 3,4 and 5) are able to classify the holdout sample well beyond chance alone. The worst performing model is Model 3 and the Cohen,s Kappa for this model is .44 and this is well beyond chance (Test of Ho: Kappa=0: z=4.40, p =0.0000 t.t.t.).

Thus, we have five models each of which is able to classify the holdout sample (n = 100) with an accuracy of between 72% (Model 3) and 78% (Model 4).

The next stage is to average the results of all models to see if this increases the accuracy over that of the highest model in the ensemble. Model 4 has an accuracy of 78% and so the ensemble will only be considered an improvement if the ensemble classifies more accurately than this.

The ensemble score is derived by averaging the logistic score for each case across the 5 models. If the average is above .5 the case is scored as a 1 while if the score is below .5 the case is scored as a 0. The result of the ensemble is a sensitivity of 82%, specificity of 78% giving an overall accuracy of 80%. The Cohen's Kappa value for this result is .6 which is significantly above chance (Test of Ho: Kappa=0: z=6.00, p =0.0000 t.t.t.). Thus, the accuracy of the ensemble of 80% is greater than the accuracy of any of the constituent models in the ensemble.

Ranking the Poems

The upshot of the preceding section is that we have an algorithm that is able to correctly classify poems as professional/amateur with an accuracy of 80% using linguistic variables. There are several applications for such an algorithm. For example, a publisher who needs a quick way of sorting through the voluminous submissions received on a weekly basis could first select a filtered list by running poems though such an algorithm. However, I wish to discuss a different application – the ranking of contemporary established poems. There is a tradition of regarding poets as "great", "minor". We tend to ignore the fact that some poets are not great or minor but are simply forgotten, as Forsythe's (2000) study emphasizes. TS Eliot points out that there is a distinction between major and minor poets but that most people would disagree about which poets should be on which lists (Eliot, 1946). The point of

ranking poems using a classification scheme such as the one advocated in this paper is that such a method provides an objective measure of the likely subjective judgments of many individuals.

The procedure is to use the ensemble classifier to give each of the established poems a score which can then be used to place them on a continuum from most professional to least professional. The score is simply the score derived by the ensemble classifier. That is, the score is the average logit score derived from the 5 logit scores of the 5 constituent models in the ensemble.

The amateur poets are excluded from this comparison for the simple reason that their status is not in contention. However it should be noted that there is no reason that we could not provide a score for the purposes of identifying amateur poems who are producing work of a professional standard. In this regard it is worthwhile noting that in the control group of 100 amateur poets, there are 22 with logit scores in the "professional" range of <.5. Of these 22, three score in the very high range of >.8 suggesting that these poems may be indicative of future poetic success.

Table 5 lists the poems and authors in descending order of logit scores. The highest score is .88 for the poem *Working Late* by Louis Simpson. The lowest score is .09 for Blackberry Eating by Galway Kinnell.

Table 5 about here

The vast majority of the poems, 86 out of 100, have scores in the "professional" range of >.5. Interestingly, 14 of the poems score in the amateur range of <.5. In other words, there are 14 poems that are more like amateur poems than professional poems. One way to explain this is that this can be expected given that the classifier has a specificity of 82%. In other

words, there will be up to 18% that are misclassified. The 14 misclassified poems represent a misclassification of 14% which is within the expected error range.

However, this interpretation has one important caveat in that when we compare the poets who have more than one poem in the corpus, there is a great deal of consistency in the classifications of their poems. Of those poets who have more than one poem in the corpus, most show consistently high or low quality. For example, Ai has two poems in the corpus, *Riot Act April 29 1992* and *Twenty Year Marriage* which score in the high to very high range of .71 and .82 respectively. At the other extreme are Galway Kinnell and Robert Creely who also have two poems each in the corpus but whose poems both score in the amateur range of <.5. Finally, there are poets who have poems in each of the high and low scoring categories. CD Wright, for example scores .47 for Approximately Forever and .78 for More Blues and the Abstract Truth. Carol frost has three poems in the corpus and these show great variation from .4 for *Sexual Jealousy* to .59 for *The Undressing* and .79 for *To Kill a Deer*. In all there are six poets who straddle the two categories. Given that there are 30 poets with more than one poem in the corpus, the majority (26) have poems in one category or another. Thus, the 6 that straddle two categories represent the exceptions rather than the norm. Furthermore, where a single poet has more than one poem in the "amateur" range, this is not merely a result of the 20% error of the classifier but may indicate that the poems are in fact more like amateur poems than professional poems.

Conclusion

In this paper I have extended the work of Kao and Jurafsky (2012) in three ways: 1) I have examined a greater number of linguistic variables and in the process I have identified a number of variables that have not previously been linked with poetic skill. Secondly I have created an ensemble classifier consisting of 5 models. The classifier has a holdout sample

accuracy of 80%. Finally, I have used the classifier to rank a corpus of contemporary American poems. This ranking is an objective means of determining which poems are more like amateur poems and which are more like professional poems.

Bibliography

Clarke, J., & Paivio, A. (2004). Extensions of the Paivio, Yuille and Madigan (1968) Norms. *Behavioral research Methods, 36*(3), 371-383.

Eliot, T. (1946). What is Minor Poetry? *The Sewanee Review, 54*(1), 1-18.

Forsythe, R. (2000). Pops & Flops: Some Properties of Famous English Poems. *Empirical Studies of the Arts, 18*(1), 49-67.

Green, E., Bodrumlu, T., & Knight, K. (2010). Automatic Analysis of Rhythmic Poetry with Applications to Generation and Translation. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 524-533). EMNLP 2010.

Hawkins, D. (2004). The Problem of Overfitting. *Journal of Chemical Information and Computer Science, 44*(1), 1-12.

Kao, J., & Jurafsky, D. (2012). A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry. *NAACL Workshop on Computational Linguistics for Literature.* Retrieved January 4, 2013, from http://www.stanford.edu/~jurafsky/kaojurafsky12.pdf

Kaplan, D., & Blei, D. (2007). A computational approach to style in American poetry. *IEEE Conference on Data Mining.* IEEE.

Paivio, A., Yuille, J., & Madigan, S. (1968). Concreteness, Imagery, and Meaningfulness Values for 925 Nouns. *Journal of Experimental Psychology, 76*(1, Pt 2), 1-25.

Pennebaker, J., Chung, C., Ireland, M., Gonzales, A., & Booth, R. (2007). The Development and Psychometric Properties of LIWC 2007. Austin, TX. Retrieved March 18, 2008, from www.LIWC.net

Pennebaker, J., Francis, M., & Booth, R. (2001). Linguistic Inquiry and Word Count (LIWC). Mahwah, NJ: Erlbaum.

Poulin, A., & Waters, M. (Eds.). (2006). *Contemporary American Poetry, eighth ed.* Houghtin Mifflin Company.

Ultsch, A., & Kämpf, D. (2004). Knowledge Discovery in DNA Microarray Data of Cancer Patients with Emergent Self Organizing Maps. *ESANN 2004 Proceedings - European Symposium on Artificial Neural Networks, 28-30 April*, (pp. 501-506). Bruges.

Watson, D., Clark, L., & Tellegen, A. (1988). Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070.

Appendix: Tables

Table 1: Parameter Estimates for Model 1.

Variable	B	Sig	Exp(B)
article	0.379	0	1.461
insight	-0.703	0.001	0.495
Constant	-1,505	0.043	0.222

Table 2: Parameter Estimates for Model 2.

Variable	B	Sig	Exp(B)
affect	-0.561	0	0.57
cogmech	-0.282	0	0.754
Constant	7	0	1177

Table 3: Parameter Estimates and Accuracy Data for Models 1 and 2

	Sensitivity	Specificity	Accuracy	Variables	Examples	B	Sig	Exp(B)
Model 1	76%	72%	74%	Article	"the", "a"	0.379	0.000	1.461
				Insight	"imagine", "contemplate"	-0.703	0.001	0.495
				Constant		-1.505	0.043	0.222
Model 2	74%	74%	74%	Affect	"gentle", "terrible"	-0.561	0.000	0.570
				Cogmech	"imagine", "consider"	-0.282	0.000	0.574
				Constant		7.000	0.000	1177

Table 4: Parameter Estimates and Accuracy Data for Models 3,4 and 5

	Sensitivity	Specificity	Accuracy	Variables	Description	B	Sig	Exp(B)
Model 3	72%	72%	72%	EMO	Emotional Content[2]	-1.305	0.000	0.271
				Constant		5.557	0.000	259.007
Model 4	78%	78%	78%	IMG	Imagery[3]	0.746	0.017	2.108
				RHY	No. of Rhyming Words[4]	-1.598	0.026	0.202
				EMOGD	Goodness deviation[5]	-2.025	0.000	0.132
				Constant		3.276	0.128	146.240
Model 5	72%	80%	76%	CON	Concreteness[6]	0.474	0.013	1.606
				GDN	Goodness[7]	-1.101	0.006	0.332
				Constant		3.182	0.196	24.096

[2] Emotional content of nouns in the 925 word sample was derived by asking subjects to rate words according to the degree to which the words would evoke a positive or negative emotional response from people. Words that elicit strong feelings get high ratings. Words that are not emotional get low ratings.

[3] Imagery of nouns in the 925 word sample was derived by asking subjects to rate words according to the degree to which it was possible to imagine an image to represent the word. Words that elicit strong/weak images get high/low ratings. Imagability is highly correlated with concreteness.

[4] The number of rhymes for words in the 925 noun sample was derived by asking subjects, for each word, whether they can think of many words that rhyme with the given word (high rating) or few words that rhyme with it (low rating).

[5] Goodness deviation was calculated by taking the absolute deviation from neutral of goodness ratings (see note 7 below).

[6] Concreteness ratings were derived by asking subjects how easy it was to form a sensory impression of the noun depicted. Those that were easy/difficult to associated with a sense were hated high/low on concreteness.

[7] Goodness ratings for nouns in the 925 noun sample were derived from subjects' impressions of the extent to which the word evokes a high level of goodness (high rating) or badness (low rating).

Table 5: Professional Poems Ranked by Logit Scores

Title	Author	Logit
Working Late	Louis Simpson	0.88
The Image	Robert Hass	0.87
How Simile Works	Albert Goldbarth	0.87
Eating Alone	LiYoung Lee	0.86
Facing It	Yusef Komunyakaa	0.86
Nostos	Louise Gluck	0.84
Hello	Naomi Shihab Nye	0.84
Twentyyear Marriage	Ai	0.82
The Room of My Life	Anne Sexton	0.82
Years End	Ellen Bryant Voigt	0.82
Dearest Reader	Michael Palmer	0.81
When You Go Away	WS Merwin	0.80
Power	Adrienne Rich	0.80
Lying in a Hammock at William Duffys Farm in Pine Island Minnesota	James Wright	0.80
University Hospital Boston	Mary Oliver	0.80
The Prediction	Mark Strand	0.80
Traveling through the Dark	William Stafford	0.79
The Small Vases from Hebron	Naomi Shihab Nye	0.79
Japan	Billy Collins	0.79
To Kill a Deer	Carol Frost	0.79
Variations On A Text	Vallejo	0.79
More Blues and the Abstract Truth	CD Wright	0.78
To Dorothy	Marvin Bell	0.78
Gin	David St John	0.78
Cleaning a Fish	Dave Smith	0.78
The Fish	Elizabeth Bishop	0.78
GlassBottom Boat	Elizabeth Spires	0.78
The Choir	Olga Broumas	0.78
Writing in the Afterlife	Billy Collins	0.77
Dream Song 172 Your face broods	John Berryman	0.77
Reuben Reuben	Michael S Harper	0.77
Fork	Charles Simic	0.77
b o d y	James Merrill	0.77
The Abduction	Stanley Kunitz	0.77
Warning to the Reader	Robert Bly	0.76
Notice What This Poem Is Not Doing	William Stafford	0.76
Crossing The Water	Sylvia Plath	0.76
Animals Are Passing From Our Lives	Philip Levine	0.76
In Trackless Woods	Richard Wilbur	0.76
Onions	William Matthews	0.75

Title	Author	Logit
Clear Night	Charles Wright	0.75
May 1968	Sharon Olds	0.75
Those Winter Sundays	Robert Hayden	0.74
at the cemetery walnut grove plantation south carolina 1989	Lucille Clifton	0.73
Charles on Fire	James Merrill	0.73
Thrall	Carolyn Kizer	0.73
Why I Am Not A Painter	Frank OHara	0.72
The Dancing	Gerald Stern	0.72
Riot Act April 29 1992	Ai	0.71
Root Cellar	Theodore Roethke	0.71
Absences	Donald Justice	0.71
The Porcelain Couple	Donald Hall	0.71
Minor Miracle	Marilyn Nelson	0.70
This Night	William Heyen	0.70
Aubade Some Peaches After Storm	Carl Phillips	0.70
Oranges	Gary Soto	0.70
The Intruder	Carolyn Kizer	0.69
Wingfoot Lake	Rita Dove	0.68
To an Adolescent Weeping Willow	Marvin Bell	0.68
They Feed They Lion	Philip Levine	0.68
Heaven as Anus	Maxine Kumin	0.67
The Strange People	Louise Erdrich	0.67
The Russian	Robert Bly	0.66
My Noiseless Entourage	Charles Simic	0.66
New Vows	Louise Erdrich	0.65
The Older Child	Kimiko Hahn	0.64
My Indigo	LiYoung Lee	0.64
Nurture	Maxine Kumin	0.64
Personal Poem	Frank OHara	0.63
Her Kind	Anne Sexton	0.61
The Stairway	Stephen Dunn	0.61
Tomatoes	Stephen Dobyns	0.60
Letter	Jean Valentine	0.59
The Undressing	Carol Frost	0.59
The Mutes	Denise Levertov	0.57
Degrees Of Gray In Philipsburg	Richard Hugo	0.57
The Summer Day	Mary Oliver	0.56
Our Lady of the Snows	Robert Hass	0.56
Audacity of the Lower Gods	Yusef Komunyakaa	0.56
Hay for the Horses	Gary Synder	0.54
A Blessing	James Wright	0.54
Adultery	James Dickey	0.53
Celestial Music	Louise Gluck	0.52
To Speak of Woe That Is in Marriage	Robert Lowell	0.52
For the Anniversary of My Death	WS Merwin	0.51

Title	Author	Logit
Fragments	Stephen Dobyns	0.51
The Singing	C K Williams	0.49
Approximately Forever	CD Wright	0.47
Scar	Lucille Clifton	0.46
The Night The Porch	Mark Strand	0.46
Dream Song 26 The glories of the world struck me	John Berryman	0.43
WeddingRing	Denise Levertov	0.41
Pacemaker	WD Snodgrass	0.41
Sexual Jealousy	Carol Frost	0.40
After Making Love we Hear Footsteps	Galway Kinnell	0.36
A Lovely Love	Gwendolyn Brooks	0.24
Playing Dead	Andrew Hudgins	0.19
The Language	Robert Creeley	0.18
The Warning	Robert Creeley	0.16
Blackberry Eating	Galway Kinnell	0.09