A Preliminary Statistical Investigation into the impact of an N-Gram Analysis Approach based on Word Syntactic Categories toward Text Author Classification*[*]

Mona Diab
Linguistics Dept. & UMIACS
University of Maryland at College Park,
Maryland 20742, USA
mdiab@umiacs.umd.edu

John Schuster
Computer Science Dept., SEAS,
George Washington University,
Washington DC 20052,USA
schuster@seas.gwu.edu

Peter Bock
Computer Science Dept. SEAS,
George Washington University,
Washington DC 20052, USA
pbock@seas.gwu.edu

Abstract

Quantitative analysis of literary style has heretofore utilized semantic elements-word counts. This research attempts to identify quantifiable syntactic elements of style that can be used for author identification. The measurement of syntactic elements utilizes a dictionary with one part of speech per word and looks at phrases delimited by punctuation marks. Different size permutations of words - referred to as grams - are counted within each text. Correlations are measured amongst the gram frequencies of eight texts pertaining to four authors, both contemporary and non-contemporary. The correlations are performed across different gram sizes of words. The same treatment is applied to a target text, the Funeral Elegy text. The approach holds for classifying texts temporally consistently across the various gram sizes. Yet a finer grained investigation is required to certify the authorship of the Funeral Elegy text.

Key words: *N-gram, Shakespeare, Middleton, Wardigo, Funeral Elegy, Author Classification*

Introduction

Literary experts refer to the *style* of the great works of literature. It is often described as a recognizable feature of an author's writings. If an expert were to be asked what makes an author's style stand out, the answer will probably fall in the realm of abstract concepts, themes and topics. Lately, a measure of literary style quantification has been the focus on semantic elements of style. By semantic elements, the reference is being made to the usage of specific adjectives or even adverbs frequently. Where very strong vocabularies are involved, and very rare words are used, intuitively, the repetition of that word would tend to indicate the same author using it across his/her writings. Often the technical term *word count* for literary works is in reference to the frequency of these specific words or specific affixes in a text. This technique involves counting every occurrence of the specified word used in a text, and comparing counts of these words among texts.

Literature being an outlet for linguistic expression can be studied from different angles. The above mentioned techniques are considered mainly in the areas of semantics and morphology. This investigation focuses on an alternative approach to the quantification of literary works. This technique is purely *syntactic* in nature. Syntax in it's own right has many levels of depth to it. Syntactic style could include things such as paragraph structure, coordinated measurements of the syntax of the first and second sentences and their relation to the third and so on. The level of sophistication possible is extremely great, but this is a search for a "quick" key. The syntactic knowledge that is used here deals with syntactic categories of words individually, i.e. a word is defined in terms of it's category be it a noun, verb…etc. The word in context - for example subject, object…etc.- is not considered at this point. The investigation abstracts away from the actual words used, reducing them to one of many predefined categories. Therefore, the proposed approach is completely oblivious to the morphological contents of the words. This point is of relevance for this investigation as it is comparing works that cut across different eras of English Literature. For the average educated English speaker, it is not too hard to discriminate between literary works pertaining to the Victorian era – Old English - and modern day works. The ability to classify the works relies mainly on morphological aspects of the words used, for example the usage of *hath* instead of *has.* Yet in this investigation the two words are treated as belonging to the same category, verb.

This paper investigates frequencies of occurrences of these word categories individually, as a single **gram**, and permutations of more than one word. A gram is defined to be a permutation of words - a sequence -

[*] ***published in Proc. of 6th International Conference on Artificial Intelligence Applications, Cairo, Egypt 1998.***

that varies in length and the words pertain to the defined syntactic categories. The length of the permutation is defined as an **n-gram,** where the n is an integer value indicating the number of words in the gram (size of the gram).

The particular event, which raised the question of whether syntactic patterns could identify a style and thus an author, was the discovery of the Funeral Elegy text. This work was only signed W. S. at the time of the investigation; the jury was still out as to the identity of its author. It has been noted as of late that it is a certified Shakespearean text, relying on both the expertise of literary experts and the semantic word count technique. With hopes of finding a clear indication one way or the other, Eight texts were chosen as the subject of the study. The investigators set out to analyze syntactic elements of William Shakespeare, Thomas Middleton – a contemporary of Shakespeare-, and Nicholas Wardigo - a 20th century playwright. By comparing the syntactic elements of their styles, eventually arriving at some interesting results that might indicate whether the Funeral Elegy was written by William Shakespeare.

This paper seeks to address whether ***an N-Gram analysis approach applied to a text provides a valid technique for text author classification.*** An **N-gram analysis** is a comparison of gram occurrences - frequencies - within a unique gram size. In order to answer this question, three experiments were performed in the sequence presented. The first experiment was concerned with the feasibility of the approach **intra-text –** comparisons performed within the same text- as well as inter-text comparisons yet restricted to comparing texts of the same author. The second experiment investigated the ability of the N-gram approach to discriminate between texts written by different authors. The third experiment aimed at finding out whether the Elegy is Shakespearean or not utilizing the N-gram technique as defined. Therefore the objective of the investigation was t**o *determine if an N-Gram analysis is a consistent and valid approach for text author classification.***

N-Gram Analysis System

Conditions & Parameters

All three experiments had the same settings. The research focused on five complete plays pertaining to two distinct eras (Victorian ca. 15th century and 20th century English literature) written by three certified different playwrights. The texts were chosen to be of diversified genres, tragedy, comedy and melodrama. The plays are as follows: *Anthony & Cleopatra* and *All's Well That Ends Well*, both written by William Shakespeare, *The Phoenix* written by Thomas Middleton, *The Bindermeyer Theory* by Nicholas Wardigo and *A Funeral Elegy for Master William Peter* of uncertified authorship. The assumption is that the chosen texts are good representatives of the works of their authors. Three of the chosen plays were divided in half and each of the halves was treated as a separate play. It was postulated that half a text is a good representation of an entire text for correlation purposes. This approach increased the number of samples present for the experiment as well as it served as a means of testing the feasibility of the approach. The eight texts' titles, abbreviation codes, author name and approximate dates written are given in the *table 1*.

Text	Abbreviation	Author	Date
All's Well that Ends Well (entire text)[1]	Allswell	William Shakespeare	16th Century
All's Well that Ends Well (first half)	AWEW1	William Shakespeare	16th Century
All's Well that Ends Well (second half)	AWEW2	William Shakespeare	16th Century
Anthony and Cleopatra (entire text)[2]	Anthcleo	William Shakespeare	16th Century
Anthony and Cleopatra (first half)	AC1	William Shakespeare	16th Century
Anthony and Cleopatra (second half)	AC2	William Shakespeare	16th Century
The Bindermeyer Theory[3]	Binder	Nicholas Wardigo	20th Century
A Funeral Elegy for Master William Peter[4]	Elegy	Uncertified	Uncertified
The Phoenix (entire text)[5]	Phoenix	Thomas Middleton	contemp. Of WS
The Phoenix (first half)	PHENX1	Thomas Middleton	contemp. Of WS

[1,2,4,5] http://the_tech.mit.edu/shakespeare/works.html

[3] http://eng.hss.cmu.edu/drama

The Phoenix (second half)	PHENX2	Thomas Middleton	contemp. Of WS

Table 1: *Titles and codes for the selected texts*

Two system parameters Gram size and word categories. were postulated. **Gram Size** is a gram of n words. Grams of different lengths were taken from the various texts. A unit increment increasing gram size was assumed to cover an adequate portion of possible syntactic category permutations in the English language. The Gram size is the experiments' *factor* as it was the variable for each treatment in the experiments. The gram sizes were in the range of a minimum of 1-Gram to a maximum of 6-Gram, where the integer indicates the length - the number of words – of a gram.

The second parameter**, word categories,** constitutes the tags assigned to the words by the implemented tagger based on their syntactic categories. Eight word categories were postulated. The chosen categories were assumed to cover as general a categorization of words of the English language as possible. The chosen categories are defined in an arbitrary order as *noun, adjective, verb, adverb, pronoun, conjunction, determiner, and preposition.* There were two main postulates associated with choice of the categories:

Postulate 1
The choice for a tag is determined manually by its usage within the context of its first occurrence.
Postulate 2
There is only one tag for a word.

Text Preprocessing & Gram Extraction

Text Tagging

The texts were preprocessed for all three experiments and the n-grams were extracted before the experiments were performed. The texts were initially broken down in **phrases**, where a phrase is defined as a sequence of words delimited by any kind of punctuation mark, beginning of a sentence or end of a sentence. Hyphens were converted into spaces. Then they were tagged based on a single entry dictionary to determine each word's category. The dictionary was manually implemented.[6] Each of the tagged phrases was put on a separate line entry creating the tagged text. After the first pass, the tags corresponding to the words in the text were converted to numerical codes based on the following formula.

$$SC_i = 2^i$$

where SC_i is a single Syntactic Category and $0 \le i \le 7$ corresponding to the Tagged Categories, and the order of the categories is arbitrary

The categories and their equivalent numerical codes are shown in *table 2*.

Tag	Code
Noun	1
Adjective	2
Verb	4
Adverb	8
Pronoun	16
Conjunction	32
Determiner	64
Preposition	128

Table 2*: Numeric Codes for each of the tagged categories*

[6] http://www.seas.gwu.edu/schuster/author_identification/diction.txt

This process was repeated for each of the eight texts.

Gram Extraction

The grams are extracted by overlapping the words within a phrase for any gram size that is greater than one. For example, if a phrase has four words , three 2-Grams are extracted as illustrated in fig.1, the codes 1, 2, 32 correspond to the categories noun, adjective and conjunction, respectively.

1	32	2	32
1	32	2	32
1	32	2	32

Fig.1 *An illustration of a 2-Gram extraction*

Within each text and gram size, the frequency of a unique gram's occurrence is tabulated. It was stipulated that as the gram size increases the gram frequency decrease due to the scarcity of long phrases in the English language. The following is an illustrated example of the 21st line from "Anthony and Cleopatra":

"News, my good lord, from Rome."

Which is broken into three phrases separated by commas. The first phrase consists of one word '*News*' which is tagged as a noun. Repeating this for '*my good lord'* is tagged as two adjectives and a noun. Finally, '*from Rome'* is tagged as preposition and noun. The numeric codes are as follows for the three phrases, each on a separate line

[1]
[2 2 1]
[128 1]

Grams of sizes 1 through 6 are extracted resulting in the following:

Six 1-Grams : {1,2, 2,1,128,1}

where code '**1'** has a frequency of three, code '**2**' has a frequency of two, and code '**128**' has a frequency of one,

Three 2-Grams: {(2 2), (2 1), (128 1)}

One 3-Gram : (2 2 1)

where each has a frequency of one.

For each factor - the different gram sizes - a table is created where the row entry is a gram permutation unified across the different texts. I.e. if a gram is present in 3 of the texts and not present in the other 5 texts, the frequencies of this gram in those five texts is 0. The order of the word categories in a gram is extremely important, i.e. the 2-gram "noun verb" is different from "verb noun" therefore constituting two different entries in the table. The data is normalized due to the disparity in the text lengths. In the grams of size 1 through 3, the permutation lists were exhaustive. For the grams of size 4 through 6 only the encountered grams were listed. Accordingly, the resulting number of grams are 8, 64, 512, 814, 495, and 85 corresponding to 1-Gram, 2-Gram, 3-Gram, 4-Gram, 5-Gram and 6-Gram, respectively. The resulting data is then utilized for the various comparisons and tests.

Experiments

The equipment, implementation and analysis environment as well as the performance metric, factor and the procedure were the same throughout the investigation. The difference lay in the combination of the texts for comparative purposes.

Performance Metric

A correlation measurement between texts within a gram size is suitable for the three experiments. The correlation was calculated between the gram frequencies of the compared texts.

Factor

The Gram sizes were in the range of 1 through 6. Greater than 6-Gram sizes - 7-Gram - were investigated but the data was very sparse as it is not highly recurrent to find phrases that are longer than 6 words in length.

Procedure

Combinations of the eight texts were compared and correlation coefficients were calculated. The correlation coefficient was calculated within a single gram size at a time, across all gram sizes. The correlation coefficient was measured based on the following Standard formula:

$$\rho_{x,y} = \frac{\operatorname{cov}(X, Y)}{\sigma_x \cdot \sigma_y}$$

where

$$-1 \leq \rho_{x,y} \leq 1$$

and

$$\operatorname{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_x)(Y_i - \mu_y)$$

X_i and Y_i correspond to frequency of a gram occurrence in texts X and Y, n is the number of grams in a specific gram size which is a constant value across the texts

The correlation coefficients were all transformed to logarithmic space in order to perform normal test statistics on the results.Shakespearean text comparison is performed by utilizing the mean correlation coefficient. Null hypotheses were asserted for experiments 1, 2 and 3. Standard hypothesis testing methods – Z tests and T student tests - were applied.

Experiment 1

The feasibility of the approach was proven with the preliminary set of tests outlined in this experiment. Intra-text comparisons were performed, texts of the same author were compared to one another. Accordingly, Shakespeare's Anthcleo was compared to his Allswell – inter-text as well as AC1 was compared to AC2, AWEW1 to AWEW2 and PHENX1 to PHENX2. Both inter-text and intra-text comparisons were performed across all gram sizes. ***Research Hypothesis 1 was that a significant correlation exists between texts by the same author.*** The following Null hypotheses were tested:

$\mathbf{H_{1.1o}}$: Correlation between AC1 and AC2 is equal to zero.
$\mathbf{H_{1,2o}}$: Correlation between AWEW1 and AWEW2 is equal to zero.
$\mathbf{H_{1.3o}}$: Correlation between Anthcleo and AWEW1 is equal to zero.
$\mathbf{H_{1.4o}}$: Correlation between PHENX1 and PHENX2 is equal to zero.

The results are tabulated in tables 3-8 in the **Results** section. The confidences obtained for the null hypotheses are tabulated in table 10.

Experiment 2

The goal of this experiment was to investigate whether the approach can serve as a tool for classifying authors based on their *syntactic* style. The analysis approach was proven to achieve this goal with the following postulated hypotheses: ***Research hypothesis 2 was that An N-Gram Analysis could successfully distinguish an Author's style.*** Two postulates and two sub-hypotheses were associated with this hypothesis. The first postulate is that an N-Gram approach is a good indication of syntactic style, and the second is that the correlation between gram frequencies within texts is a good measure of style consistency of an author. The two sub-hypotheses and their associated null hypotheses were as follows:

Hypothesis 2.1 Correlation between texts of the same author is greater than the correlation among contemporary texts of different authors

$\mathbf{H_{2.1.1o}}$: Correlation between AC2 and Allswell is less than the correlation between AC2 and Phoenix.
$\mathbf{H_{2.1.2o}}$: Correlation between PHENX1 and PHENX2 is less than the correlation between PHENX1 and Shakespearean texts.

Hypothesis 2.2 Correlation between contemporary texts is higher than the correlation between non-contemporary texts

$\mathbf{H_{2.2.1o}}$ Correlation between Shakespearean and Phoenix texts is less than the correlation between Shakespearean texts and Binder text.
$\mathbf{H_{2.2.2o}}$ Correlation between Phoenix and Shakespearean texts is less than the correlation between Phoenix and Binder text.
$\mathbf{H_{2.2.3o}}$ Correlation between Anthcleo and Allswell is less than the correlation between Anthcleo and Binder text.

Correlation coefficients are tabulated in tables 3-8. The confidences obtained for the null hypotheses are tabulated in table 10.

Experiment 3

This experiment follows experiment 2 in the sequence of the investigation. The experiment was to decide whether William Shakespeare is the author of the Funeral Elegy using the N-Gram analysis approach proposed. In this experiment the Elegy text is compared to all the other texts. ***Research hypothesis 3 is that an N-gram analysis is sufficient to indicate that Shakespeare is the author of the Funeral Elegy.*** Two sub-hypotheses are associated with this experiment. The sub-hypotheses are included with their corresponding null hypotheses.

Hypothesis 3.1 The correlation between the Elegy text and Shakespearean texts is higher than the correlation between the Elegy text and non Shakespearean texts

$\mathbf{H_{3.1.1o}}$: Correlation between Elegy and Shakespearean texts is less than the correlation between Elegy and Binder texts.
$\mathbf{H_{3.2.2o}}$: Correlation between Elegy and Shakespearean texts is less than the correlation between Elegy and Phoenix texts.

Hypothesis 3.2 The correlation between the Elegy text and Shakespearean texts is higher than the correlation between Shakespearean texts and non Shakespearean texts

$H_{3.2.1o}$: Correlation between Shakespearean and Elegy texts is less than the correlation between Shakespearean and Binder texts.
$H_{3.2.2o}$: Correlation between Shakespearean and Elegy texts is less than the correlation between Shakespearean and PHENX1 texts.

The correlation coefficients are tabulated in tables 3-8. The confidences obtained for the null hypotheses are tabulated in table 10.

Results

Correlation Coefficients for the 6 grams for the eight texts are tabulated in tables 3-8. The entries t1,t2, t3, t4, t5, t6, t7 and t8 refer to the tagged texts AC1, AC2, AWEW1, AWEW2, Binder, Elegy, PHENX1, PHENX2, respectively. 1G, 2G, 3G, 4G, 5G and 6G refer to gram sizes 1-gram, 2-gram, 3-gram, 4-gram, 5-gram and 6-gram, respectively.

t2	1.00						
t3	0.99	0.99					
t4	0.97	0.97	0.97				
t5	0.94	0.94	0.93	0.98			
t6	0.86	0.87	0.88	0.77	0.68		
t7	0.99	0.99	0.99	0.96	0.93	0.87	
t8	0.99	0.99	0.99	0.95	0.91	0.91	1.00
1-G	***t1***	***t2***	***t3***	***t4***	***t5***	***t6***	***t7***

Table 3

t2	0.99						
t3	0.98	0.98					
t4	0.97	0.97	0.97				
t5	0.92	0.90	0.89	0.95			
t6	0.87	0.87	0.89	0.70	0.81		
t7	0.98	0.98	0.99	0.91	0.97	0.88	
t8	0.98	0.98	0.99	0.89	0.97	0.90	0.99
2-G	***t1***	***t2***	***t3***	***t5***	***t4***	***t6***	***t7***

Table 4

t2	0.97						
t3	0.96	0.97					
t4	0.95	0.95	0.95				
t5	0.87	0.87	0.84	0.92			
t6	0.84	0.84	0.86	0.79	0.68		
t7	0.95	0.95	0.97	0.95	0.86	0.86	
t8	0.94	0.95	0.96	0.94	0.86	0.87	0.97
3-G	***t1***	***t2***	***t3***	***t4***	***t5***	***t6***	***t7***

Table 5

t2	0.86						
t3	0.84	0.85					
t4	0.82	0.82	0.84				
t5	0.66	0.66	0.64	0.74			
t6	0.65	0.67	0.72	0.46	0.63		
t7	0.80	0.80	0.85	0.64	0.82	0.69	
t8	0.79	0.79	0.83	0.65	0.82	0.74	0.86
4-G	***t1***	***t2***	***t3***	***t5***	***t4***	***t6***	***t7***

Table 6

t2	0.56						
t3	0.56	0.52					
t4	0.56	0.57	0.58				
t5	0.36	0.34	0.30	0.44			
t6	0.30	0.36	0.46	0.40	0.21		
t7	0.49	0.49	0.60	0.57	0.30	0.42	
t8	0.43	0.43	0.55	0.58	0.32	0.52	0.61
5-G	***t1***	***t2***	***t3***	***t4***	***t5***	***t6***	***t7***

Table 7

t2	0.25						
t3	0.28	0.16					
t4	0.30	0.20	0.29				
t5	0.20	0.10	-0.02	0.11			
t6	-0.04	-0.07	0.12	0.11	0.11		
t7	0.09	0.21	0.33	-0.01	0.44	0.15	
t8	0.18	0.07	0.25	0.07	0.33	0.30	0.37
6-G	***t1***	***t2***	***t3***	***t5***	***t4***	***t6***	***t7***

Table 8

Figures 2&3 plot relevant correlation amongst texts across the different gram sizes for comparison purposes.

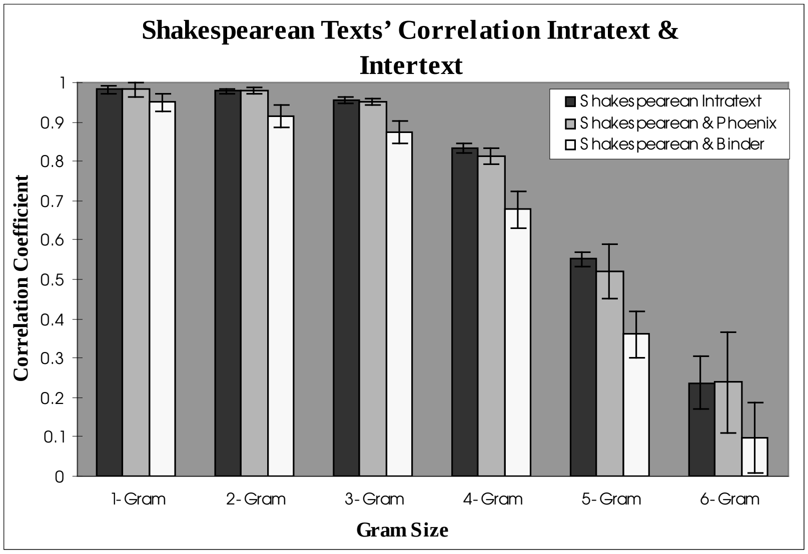

Fig. 2 Comparative correlation coefficients based on Shakespearean texts

Figure 2 depicts the correlation among Shakespearean texts - intra-text, as well as the correlation between Shakespearean texts and Phoenix text and the correlation between Shakespearean texts and Bindermeyer texts. The correlation coefficients are plotted for all gram sizes. The Shakespearean intra-text and Middleton & Shakespearean bars seem to be very close to one another, as also confirmed by the standard error.

Same plot is depicted in figure 3, where the Elegy text is compared with the Shakespearean and non-Shakespearean texts based on their corresponding correlation coefficients.

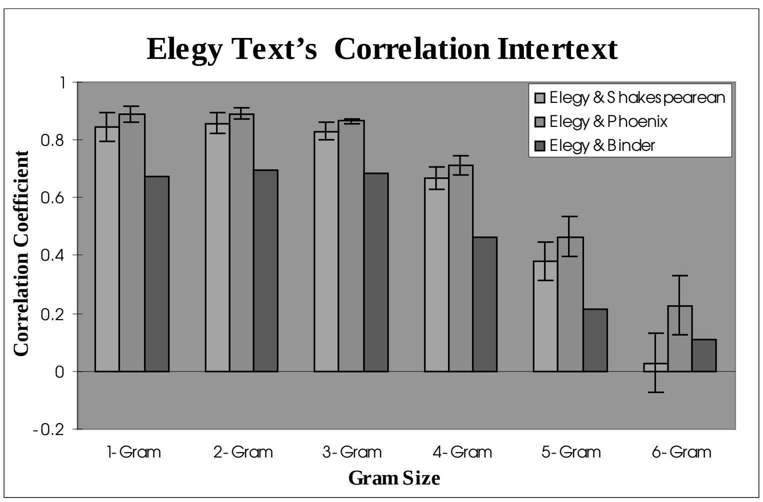

Fig.3 Comparative correlation coefficients based on the Elegy text

The Elegy text shows a higher correlation consistently, across all gram sizes, with the Shakespearean and Middleton texts than its correlation with the modern text, Binder.
In both figures 2 and 3, a downward slope is observed across the gram sizes, as the gram size increases the correlation coefficient decrease. Table 9 tabulates the results of a linear regression performed on the two plots.

Correlation	Slope Coeff.
Intra-text Shakespearean	-0.189
Shakespearean & Phoenix	-0.192
Shakespearean & Binder	-0.215
Elegy & Shakespearean	-0.211
Elegy & Phoenix	-0.173
Elegy & Binder	-0.165

***Table 9**: Regression Coefficient of the least fit line amongst correlation coefficients for various gram sizes*

The obtained confidences across the gram sizes for the null hypotheses are tabulated in table 10. The first digit in the hypothesis tag indicates the experiment number.

Null #	Null Hypotheses Statement	1-G	2-G	3-G	4-G	5-G	6-G
1.1o	Corr(AC1, AC2)=0	0.00	0.00	0.00	0.00	0.00	0.02
1.2o	Corr(AWEW1, AWEW2)=0	0.00	0.00	0.00	0.00	0.00	0.01
1.3o	Corr(Anthcleo, Allswell)=0	0.00	0.00	0.00	0.00	0.00	0.00
1.4o	Corr(PHENX1, PHENX2)=0	0.00	0.00	0.00	0.00	0.00	0.00
2.1.1o	Corr(AC2, Allswell)< Corr(AC2, Phoenix)	0.39	0.53	0.08	0.00	0.00	0.05
2.1.2o	Corr(PHENX1,PHENX2)< Corr(PHENX1, Shakespearean)	0.17	0.01	0.00	0.00	0.00	0.02
2.2.1o	Corr(Shakespearean, Phoenix) < Corr(Shakespearean, Binder)	0.13	0.00	0.00	0.00	0.00	0.00
2.2.2o	Corr(PHENX1, Shakespearean) < Corr (PHENX1, Binder)	0.07	0.00	0.00	0.00	0.00	0.00
2.2.3o	Corr(Anthcleo,Allswell) < Corr(Anthcleo,Binder)	0.15	0.00	0.00	0.00	0.00	0.00
3.1.1o	Corr(Elegy, Shakespearean) < Corr(Elegy, Binder)	0.17	0.00	0.00	0.00	0.00	1.00
3.1.2o	Corr(Elegy, Shakespearean) < Corr (Elegy, Phoenix)	0.37	0.81	0.99	0.99	1.00	1.00
3.2.1o	Corr(Shakespearean, Elegy) < Corr (Shakespearean, Binder)	0.12	0.96	1.00	0.71	0.16	1.00
3.2.2o	Corr(Shakespearean, Elegy) < Corr(Shakespearean, PHENX1)	0.02	1.00	1.00	1.00	1.00	1.00

***Table 10**: Confidences for statistically tested null hypotheses across the six gram sizes*

Conclusions & Recommendations

The results were favorable from experiment 1 and accordingly the investigation continued with experiments 2 and 3. According to confidences listed in Table 10, all the null hypotheses pertaining to Experiment 1 were rejected across the six gram sizes. The status for the alternative hypotheses for all the experiments are tabulated in table 11.

Alt. Hypo.#	Status	Confidence level
H_1	Accepted	95%
$H_{2.1}$	Accepted	89%
$H_{2.2}$	Accepted	89%
$H_{3.1}$	Judgement reserved	D.N.A
$H_{3.2}$	Accepted for 1Gram Judgement reserved rest of Grams	94% D.N.A

***Table 11**: Status of Alternative Hypotheses*

According to the listed confidences, ***hypothesis 1 is accepted with at least a confidence level of 95%*** indicating that texts belonging to the same author have a high correlation amongst them. ***Hypothesis 2 is accepted with at least an 89% confidence level***, indicating that the approach was able to distinguish the various authors apart across the different gram sizes. Yet on ***Hypothesis 3, judgement is reserved*** due to the conflicting confidences obtained. In spite of the coarseness of the analysis technique and the simplicity of the dictionary and the tagger, the N-Gram approach seems to be successful in temporal classification of the texts with no morphological reference, for instance, the words ‘have’ and ‘hath’ were both considered verbs. Looking at tables 3-8 we notice the decrease in the magnitude of the correlation coefficients as the gram size increases, yet the ratio between the coefficients seems to remain close to constant across the six gram sizes.

Interestingly enough, figure 2 seem to illustrate that the N-gram analysis approach is capable of separating out the modern non contemporary text (the Bindermeyer) from the contemporary texts. This result can be observed by looking at both the magnitudes of the correlation coefficients amongst contemporary texts in comparison to those between them and the Binder text. The standard error bars also seem to confirm the distance in the correlation coefficient magnitudes.

Looking at figure 3, the plot seems to confirm that the Elegy text is not a modern text by having a distinctly higher correlation consistently across gram sizes 1 through 5. The correlation of the Elegy with Middleton’s writings seems to be more salient than it’s correlation with the chosen Shakespearean texts. The results from the Elegy text comparisons do not add any pointers in the direction of identifying it as a work of Shakespeare. No conclusion can be drawn one way or the other due to the overlapping standard deviations of the two bars – Elegy with Shakespeare and Elegy with Middleton - across all the gram sizes,. The lack of more texts pertaining to Middleton for comparison purposes, did not allow for the investigation in that direction. Accordingly, It can be concluded that the N-gram analysis approach shows a good sign towards temporal classification as it has succeeded in classifying the Elegy text temporally yet it seems that the exact authorship requires a finer grained analysis.

Yet an intriguing observation can be made, the results from the size 1-gram consistently indicate that the Elegy is written by William Shakespeare which is consistent with the alternative approaches results that rely on semantic word count techniques.

Even though N-Gram Analysis based on simple word syntactic categories might seem to be on the opposite end of the spectrum, it yielded results that suggest the complementarity of the approach to the prevailing semantic word count technique. Accordingly, further research within this framework is recommended, hopefully with a more accurate dictionary, and a larger sample of texts, better results can be obtained. The most consistent results were obtained from grams of sizes 3, 4 and 5. The results seemed to break off at gram size 6. Further research could very possibly point towards the existence of an optimal gram size.
Also this study can be expanded further by increasing the sophistication level of the syntactic analysis both on the word and sentence levels. On the word level, syntactic categories can go a step beyond independent syntactic categories to more complex context sensitive categories such as subject, object and so forth.

Acknowledgements

The first author would like to acknowledge Taras P. Riopka (GWU) & Philip Resnik for their useful comments and acknowledge DARPA/ITO Contract N66001-97-C-8540 for travel funding as well as the department of Linguistics at University of Maryland at College Park for their support.

References

- Allen, James. *Natural Language Understanding*. Benjamin Cunnings Publishing Company, Inc., 1995.
- Bock , Peter. *The Emergence of Artificial Cognition*. World Scientific Publishing Co., 1993.
- Devore, Jay L. *Probability & Statistics for Engineering and the Sciences*. Brooks/Cole Publishing Company, 1991.
- Freund, John. *Mathematical Statistics*. Prentice Hall, 1971.
- Middleton, Thomas. *The Phoenix*. *http://the_tech.mit.edu/shakespeare/works.html*
- Shakespeare, William. *Anthony & Cleopatra*. *http://the_tech.mit.edu/shakespeare/works.html*
- Shakespeare, William. *All's Well that Ends Well*. *http://the_tech.mit.edu/shakespeare/works.html*
- Wardigo, Nicholas. *The Bindermeyer Theory*. http://eng.hss.cmu.edu/drama
- W.S. *A Funeral Elegy for Master William Peter*. http://the_tech.mit.edu/shakespeare/works.html
- Voelker, David H., Peter Z. Orton. *Cliff's Quick Review: Statistics.* Cliff Notes, 1993

Filename:	shakes1.rtf
Directory:	F:\denise\amy\treports\MONA\MONA
Template:	C:\Program Files\Microsoft Office\Office\Normal.dot
Title:	A Preliminary Statistical Investigation into the impact of using an N-Gram analysis approach based on Word syntactic categories toward Text author classification
Subject:	
Author:	mona diab
Keywords:	
Comments:	
Creation Date:	6/8/00 11:49 PM
Change Number:	2
Last Saved On:	6/8/00 11:49 PM
Last Saved By:	mona diab
Total Editing Time:	0 Minutes
Last Printed On:	6/15/00 5:41 PM
As of Last Complete Printing	
Number of Pages:	11
Number of Words:	4,252 (approx.)
Number of Characters:	24,238 (approx.)