The Thirty-Second International Florida
Artificial Intelligence Research Society Conference (FLAIRS-32)

# Learning Patterns of Assonance for Authorship Attribution of Historical Texts

**Lubomir Ivanov**
Computer Science Department, Iona College, 715 North Avenue, New Rochelle, NY 10801, U.S.A.
livanov@iona.edu

**Abstract**

This paper deals with extracting and learning patterns of assonance as a stylistic feature for author attribution of historical texts. We describe an assonance extraction algorithm and consider results from an extensive set of machine learning experiments based on a historical corpus of 18th century American and British texts. The results are compared with those obtained from the use of other prosodic and traditional stylistic features.

## 1. Introduction

Author attribution is the task of identifying the writer of a text, whose authorship is either unknown or disputed. Historically, attribution has been performed by humanities experts, but human attribution is time consuming and prone to errors and subjective opinions. The advances in natural language processing, machine learning, and data mining have led to a significant interest in automated attribution, which tends to be more objective, thorough, and capable of uncovering inconspicuous stylistic subtleties. Automated attribution has been used to re-examine the authorship of many literary works throughout the ages (Hoover 2018, Jackson 2014, Burrows 2012, Craig and Kinney 2009, Jackson 2008, Binongo 2003, Matthews 1995, Matthews and Merriam 1993, Holmes 1992, Barquist and Shie 1991, Burrows 1987, Lowe and Smith 1985, Morton 1965) and to determine the authorship of historically significant documents (Mosteller and Wallace 1964, Petrovic et al 2016, Petrovic et al 2015, Petrovic et al 2014). Modern applications of authorship attribution include digital copyright- and plagiarism detection, gender identification, forensic linguistics, and criminal/anti-terror investigation. (Ogaltsov and Romanov 2017, Tellez et al 2017, Agrawal and Gonçalves 2016, Garciarena et al 2016, Kuznezov et al 2016, Sausa-Silva 2012, Zheng et al 2006, Kotzé 2005, Abbasi and Chen 2005, Argamon et al 2003, de Vel et al 2001).

Authorship attribution is based on selecting and learning an appropriate set of stylistic features, which capture the intuitive notion of an author's style. Such features have traditionally included function words, character-/word-n-grams, part-of-speech (PoS) tags, sentence lengths, etc. The frequencies of use of these features in known works are used to train classifiers to recognize each author's writing style. Once trained, the models can be applied to recognizing the style of a document of unknown or disputed authorship. For an overview of the field, the reader is referred to these surveys (Stamatatos 2016, Stamatatos 2009, Juola 2008)

Recently, there has been interest in using prosodic features as stylistic markers for authorship. Prosody defines how speech elements larger than phonemes are articulated. Prosodic features such as intonation, stress, and tempo and prosodic poetic devices such as alliteration, assonance, and consonance often carry an emotive charge, and can be used by an author to emphasize a particular point. Many 18th century authors were well aware that their works will be read out in public and made careful use of prosody in their writing. Thus, learning unique prosodic patterns from historical texts may provide an alternative approach to historical attribution. Lexical stress for authorship attribution was considered in (Dumalus and Fernandez 2011) and explored in depth in (Ivanov et al 2018, Ivanov and Petrovic 2015). In (Ivanov 2016), the role of alliteration for author attribution was investigated. Both lexical stress and alliteration appear to be moderately successful when the author pool is small.

In this paper, we explore the usefulness for author attribution of another prosodic feature – assonance. We present an algorithm for extracting assonance from text and describe a set of assonance-based, machine learning experiments based on a historical corpus of 18th-century writings. This corpus is used to compare the performance of assonance to the that of traditional stylistic features, lexical stress, and alliteration. The results of combining assonance with other stylistic features using an ensemble of classifiers are also presented.

## 2. Assonance

### 2.1 Definition

Assonance is a literary technique defined as the use of a repeated vowel or diphthong sound in nearby (often non-rhyming) words. Examples of assonance abound in literature:

- "Tyger, Tyger burning bright in the forest of the night" (W. Blake)
- "I wandered lonely as a cloud
  That floats on high o'er vales and hills,
  When all at once I saw a crowd,
  A host of golden daffodils;" (W. Wordsworth)

Unlike alliteration, which is usually easy to recognize, assonance is far more subtle: The repeated nearby vowel sounds echo off each other and create a mood, which affects the reader/listener subconsciously. There has been little work on assonance except for a recent paper (Addanki and Wu 2013) on rhyme identification in hip hop music. A few earlier works (Genzel et al 2010, Byrd and Chodorow 1985) also focus on rhyme identification, including briefly touching on assonance. To the best of our knowledge, assonance has not been used in authorship attribution studies.

As with most literary terms, there is no precise definition of what constitutes assonance: The phrase "in nearby words" in the definition of assonance does not specify anything about how close assonant sounds must be. Thus, we consider the inter-vowel distance as variable, and experimented with multiple distance values.

### 2.2 Extracting Assonance from Text

Our assonance extraction algorithm uses a modified version of the popular Carnegie Mellon University (CMU) pronunciation dictionary. The CMU dictionary contains 133854 word-pronunciation pairs, which have been augmented with an additional set of 1861 word-pronunciation pairs extracted from our historical corpus. The proper pronunciations of the additional historical words have been confirmed by experts in 18th century American and English literature. The dictionary specifies word pronunciations based on 39 phonemes with the vowels marked with 0 (no stress), 1 (primary stress), or 2 (secondary stress). The assonance algorithm takes as input a text and several user-specified parameters:

- Maximum distance between assonant vowels
- Search scope
- Primary-stressed-vowel or any-vowel assonance
- Longest- or longest-two assonance sequences

The first parameter defines how far apart vowels or diphthongs can be to be considered assonant. The second parameter specifies the scope for the assonance search: within sentences, within paragraphs, or throughout the whole text. The third parameter indicates whether only primary stressed vowels or any vowels should be considered. The fourth parameter indicates whether only the longest or the two longest assonance patterns per block (sentence, paragraph, text) should be considered. In the latter case, the secondary pattern must be at least 80% of the length of the primary assonance pattern.

A pseudocode version of the algorithm is presented below:

```
Create an AssonanceMap for <assonanceLabel, count> pairs;
for (each wordblock WB)  {
    Form a new string "vowelBlockString"
    for (each word W in WB)
       Add to vowelBlockString the vowels-only pronunciation of W extracted from CMUDictionary
    Create a new vowelMap for storing <vowelString, <assonanceCount, vowelsApart>> pairs;
    Set numberOfVowelSounds = 0;
    for (each vowel V in vowelBlockString){
        numberOfVowelSounds++;
        if (vowelMap does not contains a key V)
           Add V to vowelMap with assonanceCount=1 and vowelsApart =0
        else
           Increment the assonanceCount of V and set vowelsApart=0
        for (each K different from V in vowelMap)
           Increment the vowelsApart value for K
    }
    Find and label the longest (and second longest) assonance sequence(s) with their vowel sound(s) and:
    - s (short):  seqLength <= .25*numberOfVowelSounds
    - m (medium):
        seqLength  <= .5*numberOfVowelSounds AND
        seqLength > .25*numberOfVowelSounds
    - l (long):
        seqLength  <= .75*numberOfVowelSounds AND
        seqLength  > .5*numberOfVowelSounds
    - vl (very-long):
        seqLength > .75*numberOfVowelSounds
    If the above label(s) are not in AssonanceMap, add the label(s) with count=1 (e.g. <"AA_s", 1>), otherwise, increment the label's count in AssonanceMap
}
for (each key L in AssonanceMap.keySet()) {
    Compute the frequency of label L
    Write "L: frequency_of_L" to the output file
}
```

The algorithm begins by replacing each word in the text with its pronunciations from the CMU dictionary. All consonants are stripped, while vowels/diphthongs and punctuation are preserved. Next, the text is broken into blocks based on the user-specified scope (sentence/paragraph/full-text). In every block, the algorithm examines each vowel/diphthong: If the vowel has not been seen before, it is added to a vowel map with a *count* value of 1 and *vowelsApart* value set to 0. Otherwise the vowel's count is incremented and *vowelsApart* is reset. The algorithm then iterates through all map entries, incrementing *vowelsApart* for all map entries different from the selected vowel. At the end of the block, either the longest or the two longest assonance sequences are selected (based on the user-specified command-line parameter) and labeled with the vowel/diphthong they represent plus a short (s), medium (m), long (l), or very-long (vl)

tag, e.g. "AE_vl". A short assonance sequence is one consisting of less than 25% of the vowels in the block, a medium sequence is between 25% and 50%, a long sequence is between 50% and 75%, and a very long sequence is more than 75% of the length of the block. The label(s) are entered into an assonance map, which tracks the number of times different assonance sequences are encountered in the text. In the end, the entries in this map and their frequencies are output to a file. The algorithm was implemented in a multithreaded program, which executes as a pool of up 6 threads to speed up processing time. When all files have been processed, a separate program creates the training/testing vectors from all output files and writes them to an ARFF file for the WEKA data mining software (Hall et al, 2009).

## 3. Assonance Experiments

### 3.1 The Historical Document Corpus

The historical corpus used in our experiments consists of 224 attributed English language documents created by 38 authors during the second half of 18th Century (Table 1). The attribution of the documents is fairly certain, but there are several corpus-related issues - the small per-author document sets, the varying number of documents per author, the unequal document lengths, and occasional OCR errors.

| Authors | # of Texts |
|---|---|
| John Adams | 10 |
| Joel Barlow | 4 |
| Anthony Benezet | 5 |
| James Boswell | 5 |
| James Burgh | 7 |
| Edmund Burke | 6 |
| Charles Carroll | 3 |
| John Cartwright | 13 |
| Cassandra (pseud. of J. Cannon) | 4 |
| Earl of Chatham (W. Pitt Sr.) | 3 |
| John Dickinson | 4 |
| Philip Francis | 4 |
| Benjamin Franklin | 9 |
| George Grenville | 3 |
| Samuel Hopkins | 5 |
| Francis Hopkinson | 21 |
| Thomas Jefferson | 7 |
| Marquis de Lafayette | 5 |
| Thomas Macaulay | 7 |
| James Mackintosh | 7 |
| William Moore | 5 |
| William Ogilvie | 4 |
| Thomas Paine | 11 |
| Richard Price | 4 |
| Joseph Priestley | 5 |
| Benjamin Rush | 6 |
| George Sackville | 2 |
| Granville Sharp | 8 |
| Earl of Shelburne (William Petty) | 3 |
| Thomas Spence | 6 |
| Charles Stanhope | 2 |
| Sir Richard Temple | 2 |
| John Horne Tooke | 4 |
| John Wesley | 4 |
| John Wilkes | 5 |
| John Witherspoon | 8 |
| Mary Wollstonecraft | 7 |
| John Woolman | 6 |

*Table 1: Authors of attributed historical documents*

### 3.2 Baseline Experiments

The baseline experiments were conducted using the JGAAP authorship attribution software (Juola, 2009). We used the full set of historical documents and random subsets of 15, 10, and 7 authors. The stylistic features used are described in Table 2. The classification was performed with WEKA support vector machines with sequential minimal optimization (SMO) and with multilayer perceptrons (MLP)[1]. The results are summarized in Table 2 below:

| # of Authors / Classifier/ Stylistic Feature | 38 | 15 | 10 | 7 |
|---|---|---|---|---|
| *MLP/Function Words* | 67.86% | 85.86% | 90.16% | 92.31% |
| *MLP/Char-2-Grams* | 70.09% | 80/81% | 82.25% | 87.18% |
| *MLP/FirstWordInSent* | 41.52% | 64.65% | 80.33% | 89.74% |
| *MLP/Prepositions* | 58.04% | 69.70% | 91.80% | 97.44% |
| *MLP/Suffices* | 58.93% | 76.77% | 86.89% | 84.62% |
| *MLP/VowelInitWords* | 64.73% | 81.82% | 86.72% | 89.74% |
| *SMO/FunctionWords* | 68.75% | 85.86% | 85.25% | 92.31% |
| *SMO/Char-2-Grams* | 61.16% | 81.82% | 90.16% | 92.31% |
| *SMO/FirstWordInSent* | 37.05% | 65.66% | 91.80% | 97.44% |
| *SMO/Prepositions* | 54.46% | 79.80% | 88.52% | 89.74% |
| *SMO/Suffices* | 56.25% | 74.75% | 85.25% | 92.31% |
| *SMO/VowelInitWords* | 60.27% | 84.85% | 96.72% | 94.87% |
| ***AVERAGE (MLP):*** | **60.20%** | **75.76%** | **86.36%** | **90.17%** |
| ***AVERAGE (SMO):*** | **56.32%** | **78.79%** | **89.62%** | **93.16%** |
| ***Overall AVERAGE:*** | **58.26%** | **77.41%** | **87.99%** | **91.67%** |

*Table 2: Baseline accuracies (historical corpus)*

### 3.3 Assonance Experiments

We conducted a large number of experiments based on the historical corpus described in Table 1, varying the parameters of the assonance extraction program. In all experiments we used leave-one-out (L1O) validation. The first set of 18 experiments was conducted with the full set of 38 authors/224 documents using both sentence and paragraph boundaries, the

[1] Both WEKA SMO and WEKA MLP are standard models in JGAAP.

longest assonance sequences only, and an all-vowel (stressed and non-stressed) option. The inter-vowel distance was set to 5, 10, and 15 respectively. Three different WEKA classification methods were used – MLP, SMO, and Random Forest (RF). The maximum accuracy obtained in all experiments 29.91%, the average accuracy was 27.21%. The results were consistent across all experiments, with standard deviation of 1.84%. The maximum accuracy was obtained in the experiment using an MLP classifier, an inter-vowel distance of 15, and a sentence boundary. The next set of 18 experiments involved the pair of longest assonance sequences per block with all other parameters kept the same. The maximum accuracy achieved was 31.25%, and the average accuracy - 30.51%. The maximum accuracy was obtained using a SMO, an inter-vowel distance of 15, and a sentence boundary.

On the full set of 38 authors/224 historical documents, and with an average accuracy around 30%, assonance compares unfavorably with most baseline stylistic features. Its performance is closest to the first-word-in-sentence stylistic feature and is significantly lower than the top performing MW function words and character-2-gram features. A close examination of the individual author results reveals that assonance works well with some writers (e.g. Stanhope: f-measure: 1.00, Hopkins: f-measure: 0.889, Hopkinson: f-measure: 0.766), but fails with others (e.g. Barlow, Wesley: f-measure: 0.00). The set of authors for whom assonance works well is relatively consistent regardless of the choice of the algorithm parameters. Thus, we conjecture that the authors for whom assonance yields strong attribution results are those that actively use assonance in their work. Moreover, these writers have (consciously or unconsciously) uniquely integrated assonance into their writing style, and the classifiers can correctly associate specific assonance sequences and their lengths with these authors. But the classifiers have difficulties with authors, who do not use assonance or use it indistinctly. To test this conjecture, we selected the set of 15 authors, who consistently yielded a high f-measure in all attribution experiments and conducted another set of 36 experiments based on the algorithm parameters used in the all-documents experiments. This time, the average accuracy obtained was 58.42% and the maximum accuracy was 63.27%. This is close to the performance of some standard JGAAP methods (first-word-in-sentence and prepositions), though still significantly lower than MW function words and character-2-grams.

Limiting the author set to the 10 writers used in the baseline experiments produced a 73.77% accuracy. The baseline 7-author set produced an 82.98% average accuracy with assonance, while the set of 7 top-performing (i.e. highest f-measure) authors yielded an accuracy of 94.59%. These results indicate that assonance can potentially be used as a stylistic feature, provided that the candidate authors set is relatively small. Since, the number of authors in most of our actual attribution experiments is usually between 4 and 13, assonance may prove to be helpful in determining the true authorship of historically significant documents and texts.

### 3.4 Comparison to Lexical Stress and Alliteration

It was interesting to compare the performance of assonance with lexical stress and alliteration. Table 3 lists the results:

| Stylistic feature / # Authors / Learning Method | Lex.Stress (PoS) | Alliteration | Assonance |
|---|---|---|---|
| *38 authors/SMO* | 39.46% | 27.93% | 31.25% |
| *38 authors/MLP* | 47.53% | 20.57% | 30.80% |
| *13 authors/SMO* | 69.41% | 56.70% | 66.67% |
| *13 authors/MLP* | 73.56% | 53.61% | 66.67% |
| *7 authors/SMO* | 91.09% | 83.33% | 91.89% |
| *7 authors/MLP* | 90.00% | 80.93% | 94.59% |

*Table 3: Comparison between the accuracy of PoS-based lexical stress, alliteration, and assonance*

The results indicate that, in terms of accuracy, assonance falls between lexical stress and alliteration. For small sets of authors, particularly those using assonance in their writings, employing assonance as a style discriminator yields very strong results – on par with the top performing stylistic features. However, for a randomly selected set of authors, assonance is generally weaker than lexical stress but consistently stronger than alliteration in all experiments.

## 4. Assonance Combined with Other Features

Previous work (Poulston 2017, Petrovic 2016) has demonstrated that ensemble classifiers, which collectively perform author selection based on a weighted average of their individual predictions, outperform individual stylistic-feature/classifier pairs. We wanted to see how well such an ensemble will perform without assonance (as a baseline), and with assonance added to the ensemble. To set up an ensemble classifier, we paired each stylistic feature with a SMO or an MLP classifier and computed the individual L1O accuracies of all feature/classifier pairs. We then set each pair's candidate author support to be proportional to the pair's L1O accuracy. Pairs with L1O accuracy lower than the median are excluded. The overall support for each author is calculated by adding the supports the author received from the feature/classifier pairs that selected him/her.

| Stylistic features / # Authors | Traditional Features Only (Baseline) | Traditional Features + Assonance |
|---|---|---|
| *38 authors* | 77.23% | 77.23% |
| *10 authors* | 92.21% | 95.11% |
| *7 authors* | 94.87% | 96.47% |

*Table 4: Ensemble performance (without and with assonance)*

We conducted six sets of ensemble classifier experiments – with the full set of 38 authors, and with the previously used, randomly selected set of 10 and 7 baseline authors. In the first three experiments, we estimated the performance of the ensemble classifier using only the baseline feature/classifier pairs. Next, we added assonance, paired both with a SMO and an MLP classifier. The parameters used in the experiments were as follows: an inter-vowel distance of 20, a paragraph boundary, longest two assonance sequences per block, and any-vowel assonance. We repeated the 38-, 10- and 7-author experiments. The results are shown in Table 4. Assonance has no effect on the classification accuracy of the ensemble when the full set of authors is used because it exhibits a lower-than-the-median L1O accuracy and is eliminated from the decision making. However, for 10 authors, assonance clearly has an effect on the ensemble's classification accuracy, raising the overall accuracy by as much as 3%. A closer examination of the individual author predictions indicates that, in some cases where other methods failed to make a strong prediction, assonance cast the deciding vote by having a stronger maximum support and a more focused probability distribution than the other methods. The 7-author experiments confirmed these results and findings. Thus, assonance appears to be most useful as a second-level attribution feature: If a large number of candidate authors is present, traditional features such as MW function words, and character-/word-n-grams can narrow the pool of potential authors to a smaller subset, where additional features such as assonance can fine-tune the attribution prediction.

## 5. Conclusion and Future Work

In this paper we presented an experimental study of the effectiveness of the assonance prosodic feature as a stylistic marker in authorship attribution experiments. Given the research interests of our humanities colleagues, our primary focus has been on exploring the usefulness of assonance in attribution of 18th century historical texts and documents. However, we also experimented with two small poetry corpora, producing some interesting observations. Additionally, we experimented with the widely used Reuters corpus (NIST). The Reuters tests yielded similar results to those obtained in the historical corpus experiments. We also demonstrated that adding assonance to an ensemble classifier can improve the results and provide a stronger hypothesis as a starting point for humanities researchers to explore.

The results of using assonance for attributing poetry stimulated our interest in further exploring authorship attribution of poetry. We intend to use the literary resources of Project Guttenberg (Guttenberg) and other online sources to construct an appropriately-sized poetic corpus, stratified by time periods, to which prosody-based authorship attribution can be applied. While there has been sporadic work on poetry attribution (Hoover 2005, Al-falahni et al 2015, Reza 2008), we have yet to come across a truly comprehensive study of authorship attribution in poetry with its unique challenges. We intend to carry out a thorough investigation of attribution in poetry, its sub-types, and its evolution through the ages.

Finally, we are aware of the low statistical significance of many of the results based on our small 18th century historical corpus. In an effort obtain a much stronger statistical basis for our studies, we are working with colleagues in the humanities to construct a larger corpus of several thousand 18th century American newspaper articles focused on the political, social, and economic events of the day. The newspaper corpus will provide us with a new set of historical texts for our prosody-based authorship attribution studies.

## Acknowledgements

This work was supported by a grant from the R. D. L. Gardiner foundation, to whom we express our appreciation.

## References

Abbasi, A., Chen, H. 2005. Applying authorship analysis to extremist group web forum messages. *IEEE Intelligent Systems*, 20(5), pp. 67-75

Addanki K., Wu D. 2013. Unsupervised Rhyme Scheme Identification in Hip Hop Lyrics Using Hidden Markov Models. In: Dediu AH., et al, (eds) Statistical Language and Speech Processing. *SLSP 2013. Lecture Notes in Computer Science*, vol. 7978. Springer, Berlin, Heidelberg

Agrawal M., Gonçalves T. 2016. Age and Gender Identification using Stacking for Classification—Notebook for PAN at CLEF 2016. In Balog, et al, (eds), *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, Évora, Portugal, 9/16. CEUR-WS.org. ISSN 1613-0073.

Al-falahni, A., Bellafkin, M., Romdani, M., Al-Serem, M. 2015. Authorship Attribution of Arabic Poetry, *10th International Conference on Intelligent Systems: Theories and Applications (SITA)*, **DOI:** 10.1109/SITA.2015.7358411

Argamon, S., Saric, M., Stein, S. 2003. Style mining of electronic messages for multiple author- ship discrimination, *Proceedings of the 9th ACM SIGKDD*, pp. 475-480

Barquist, C., Shie, D. 1991. Computer Analysis of Alliteration in Beowulf Using Distinctive Feature Theory. *Literary and Linguist Computing* 6 (4): pp. 274-280. doi: 10.1093/llc/6.4.274

Binongo, J. N. G. 2003. Who wrote the 15th book of Oz? An application of multivariate statistics to authorship attribution. *Computational Linguistics* 16(2)

Byrd, Roy J. and Chodorow, Martin S. 1985. Using an online dictionary to find rhyming words and pronunciations for unknown words. In *Proceedings of ACL*.

Burrows, J. 2012. A second opinion on 'Shakespeare and authorship studies in the twenty-first century'. *Shakespeare Quarterly*, **63**(3): 355–92.

Burrows, J. 1987. Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method. *Oxford: Clarendon Press*

Craig, H., and Kinney, A., eds. 2009. Shakespeare, Computers, and the Mystery of Authorship. *Cambridge University Press*.

deVel,O., Anderson, A., Corney, M., Mohay, G. M. 2001. Mining e-mail content for author identification forensics. *SIGMOD Record*, 30(4), pp. 55-64

Dumalus, A., Fernandez, P. 2011. Authorship Attribution Using Writer's Rhythm Based on Lexical Stress, *11th Philippine Computing Science Congress*, Naga City, Philippines

Garciarena Ucelay J, Villegas P., Funez D., Cagnina L., Errecalde M., Ramírez-de-la-Rosa G., and Villatoro-Tello E. 2016. Profile-based Approach for Age and Gender Identification—Notebook for PAN at CLEF 2016. In Balog, et al, (eds), *CLEF 2016 Eval. Labs and Workshop – Working Notes Papers*, Évora, Portugal, 9/16. CEUR-WS.org. ISSN 1613-0073.

Genzel, Dmitriy, Uszkoreit, Jakob, and Och, Franz. 2010. "Poetic" statistical machine translation: Rhyme and meter. In *EMNLP*.

Guttenberg: https://web.eecs.umich.edu/~lahiri/gutenberg_dataset.html

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. 2009. The WEKA Data Mining Software: An Update; *SIGKDD Explorations*, Volume 11, Issue 1

Holmes, D. I. 1992. A Stylometric Analysis of Mormon Scripture and Related Texts. *Journal of the Royal Statistical Society Series A*: 155(1)91-120

Hoover, D., 2017, Authorship Attribution Variables and Victorian Drama: Words, Word-Ngrams, and Character-Ngrams, *Proc. of the 2018 Digital Humanities Conference*, Mexico City, Mexico, 6/18, pp. 212-214,

Hoover, D. 2005. Delta, Delta Prime, and Modern American Poetry: Authorship Attribution Theory and Method. In *Proceedings of 2005 ALLC/ACH Conference*, 79-80, Victoria, Canada.

Ivanov L., Aebig A., Meerman S. 2018. Lexical Stress-Based Authorship Attribution with Accurate Pronunciation Patterns Selection: *21st International TSD Conference 2018*, Brno, Czech Republic, 9/18, pp. 67-75. 10.1007/978-3-030-00794-2_7.

Ivanov. L. 2016. Using Alliteration in Authorship Attribution of Historical Texts, *Text, Speech, and Dialogue. TSD 2016. Lecture Notes in Computer Science*, vol. 9924. Springer

Ivanov, L., Petrovic, S. 2015. Using Lexical Stress in Authorship Attribution of Historical Texts, Chapter, *Lecture Notes in Computer Science*: TSD, v.9302, pp.105- 113

Jackson, MacD. 2014. Determining the Shakespeare Canon: Arden of Faversham and A Lover's Complaint. *Oxford U. Press*.

Jackson, MacD. 2008. New research on the dramatic canon of Thomas Kyd. *Research Opportunities in Medieval & Renaissance Drama*, 47: 107-127.

Juola, P. 2009. JGAAP: A system for comparative evaluation of authorship attribution. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, 1(1): 1-5.

Juola, P. 2008. Authorship Attribution. *Foundations and Trends in Information Retrieval 1*, pp. 234– 334

Kotzé E. 2010. Author identification from opposing perspectives in forensic linguistics. *Southern African Linguistics and Applied Language Studies*, 28(2): pp. 185-197

Kuznetsov M., Motrenko A., Kuznetsova R., and Strijov V. 2016. Methods for Intrinsic Plagiarism Detection and Author Diarization—Notebook for PAN at CLEF 2016. In Balog, et al, (eds), *CLEF 2016 Evaluation Labs and Workshop*, Évora, Portugal, 9/16. CEUR-WS.org. ISSN 1613-0073.

Lowe, D. and Matthews: R. 1995. Shakespeare vs. Fletcher: A Stylometric Analysis by Radial Basis Functions. *Computers and the Humanities*, 29, pp. 449-461

Matthews, R., and T. Merriam. 1993. Neural computation in stylometry: An application to the works of Shakespeare and Fletcher. *Literary and Linguistic Computing*. 8.4, pp. 203-209

Morton, A.Q. 1965. The Authorship of Greek Prose. *Journal of the Royal Statistical Society*, 128, pp. 169-233

Mosteller, F, Wallace, D. 1964. Inference and disputed authorship: The Federalist, Reading, MA: *Addison-Wesley*

NIST: https://trec.nist.gov/data/reuters/reuters.html

Ogaltsov A. and Romanov A. 2017. Language Variety and Gender Classification for Author Profiling in PAN 2017—Notebook for PAN at CLEF 2017. In Cappellato, et al, (eds), *CLEF 2017 Eval. Labs and Workshop – Working Notes Papers*, Dublin, Ireland, 9/17. CEUR-WS.org. ISSN 1613-0073

Petrovic, S., Berton, G., Campbell, S., Ivanov, L. 2015. Attribution of 18th Century Political Writings Using Machine Learning. *Journal of Technologies in Society*, v.11, iss. 3, pp. 1-13

Petrovic, S., Berton, G., Schiaffino, R., Ivanov, L. 2016. Examining the Thomas Paine Corpus: Automated Computer Author Attribution Methodology Applied to Thomas Paine's Writings. *Chapter, New Directions in Thomas Paine Studies, Edition: 1*, Publisher: Palgrave Macmillan US, Editors: S. Cleary I. Stabell, DOI: 10.1057/9781137589996

Petrovic, S., Berton, G., Schiaffino, R., Ivanov, L. 2014. Authorship Attribution of Thomas Paine Works. *International Conference on Data Mining DMIN'14*. pp. 182-188. Springer

Poulston A., Waseem Z., Stevenson M. 2017. Using TF-IDF n-gram and Word Embedding Cluster Ensembles for Author Profiling—Notebook for PAN at CLEF 2017 In Cappellato, et al, (eds) *CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers,* 9/17, Dublin, Ireland, CEUR-WS.org. ISSN1613-0073.

Raza, A, Athar, A., Nadeem, S. 2008. N-Gram Based Attribution in Urdu Poetry, *Proc. of Conf. on Language & Technology*, 88-93

Smith, M.W.A. 1985. An Investigation of Morton's Method to Distinguish Elizabethan Playwrights. *Computers & the Humanities.*, 19, 3-21

Sousa-Silva R. 2016. Detecting Translingual Plagiarism: A Forensic Linguistic Contribution to Computational Processing: http://www.uniweimar.de/medien/webis/events/pan-16

Stamatatos E. 2016. Authorship Verification: A Review of Recent Advances *Research in Computer Science*, 123, pp. 9-25, IPN.

Stamatatos E. 2009. E.A Survey of Modern Authorship Attribution Methods Journal of the American Society for Information Science and Technology, 60(3), pp. 538-556, Wiley.

Tellez E., Miranda-Jiménez S., Graff M., and Moctezuma D. 2017. Gender and language-variety Identification with MicroTC—Notebook for PAN at CLEF 2017. In Cappellato, et al., (eds), *CLEF 2017 Eval. Labs and Workshop*, Dublin, Ireland, 9/17. CEUR-WS.org. ISSN 1613-0073.

Zheng, R., Li, J., Chen, H., Huang, Z. 2006. A framework for authorship identification of on- line messages: writing style features and classification techniques. *Journal of the American Society of Information Science and Technology*, 57(3), pp. 378-393