An Investigation of Keystroke and Stylometry Traits for Authenticating Online Test Takers

John C. Stewart, John V. Monaco, Sung-Hyuk Cha, and Charles C. Tappert,
Seidenberg School of CSIS, Pace University, White Plains, NY 10606, USA

Abstract

The 2008 federal Higher Education Opportunity Act requires institutions of higher learning to make greater access control efforts for the purposes of assuring that students of record are those actually accessing the systems and taking exams in online courses by adopting identification technologies as they become more ubiquitous. To meet these needs, keystroke and stylometry biometrics were investigated towards developing a robust system to authenticate (verify) online test takers. Performance statistics on keystroke, stylometry, and combined keystroke-stylometry systems were obtained on data from 40 test-taking students enrolled in a university course. The best equal-error-rate performance on the keystroke system was 0.5% which is an improvement over earlier reported results on this system. The performance of the stylometry system, however, was rather poor and did not boost the performance of the keystroke system, indicating that stylometry is not suitable for text lengths of short-answer tests unless the features can be substantially improved, at least for the method employed.

1. Introduction

The main application of interest in this study is verifying the identity of students taking online tests, an application that is becoming more important with the student enrollment of online classes increasing, and instructors and administrations becoming concerned about evaluation security and academic integrity. The 2008 federal Higher Education Opportunity Act (HEOA) requires institutions of higher learning to make greater access control efforts for the purposes of assuring that students of record are those actually accessing the systems and taking exams in online courses by adopting identification technologies as they become more ubiquitous [8]. To meet the needs of this act, the keystroke biometric seems appropriate for the student authentication process. Stylometry appears to be a useful addition to the process because the correct student may be keying in the test answers but a coach could be providing the answers and the student merely typing the coach's words without bothering to convert the linguistic style into his own.

Keystroke biometric systems measure typing characteristics believed to be unique to an individual and difficult to duplicate [5, 9]. The keystroke biometric is a behavioral biometric, and most of the systems developed previously have been experimental in nature. Nevertheless, there has been a long history of commercially unsuccessful implementations aimed at continuous recognition of a typer. More recently several commercial products have been developed for hardening passwords in computer security schemes [1, 4]. While most previous work dealt with passwords or short name strings [2, 6, 14, 18, 19], some used long-text input [3, 7, 12, 22].

Stylometry is the study of determining authorship from the linguistic styles of the authors [21]. Traditionally, it has been used to attribute authorship to anonymous or disputed literary documents. More recently, computer-based communication and digital documents have been the focus of research [10, 20], sometimes with the goal of identifying perpetrators or other malicious behavior. Recent computer studies have used stylometry to determine authorship of emails [16, 27] and tweets [11] as efforts to authenticate uses of more common digital media.

The keystroke and stylometry biometrics are appealing for this application for several reasons. First, they are not intrusive to computer users. Second, they are inexpensive since the only hardware required is a computer with keyboard. Third, text continues to be entered for potential repeated checking after an initial authentication phase, and this continuing verification throughout a computer session is referred to as dynamic verification [12, 13].

A number of measurements or features are generally used to characterize an individual. For the keystroke biometric these measurements are typically key press duration (dwell) times, transition (latency) times, and the identity of the keys pressed. Stylometry typically uses statistical linguistic features at the word and syntax level.

The current work extends our prior studies on a robust keystroke biometric system for long-text input [22, 23, 26]. This system is unique in several respects. First, it can collect raw keystroke data over the Internet as well as from a key logger on an individual machine. Second, it focuses on long-text input where sufficient keystroke data are available to permit the use of powerful statistical feature

978-1-4577-1359-0/11/$26.00 ©2011 IEEE

measurements – and the number, variety, and strength of the measurements used in the system are much greater than those used by earlier systems reported in the literature. Third, it focuses on applications using arbitrary text input because copy texts are unacceptable for most applications of interest. And, fourth, because of the statistical nature of the features and the use of arbitrary text input, special statistical fallback procedures are incorporated into the system to handle the paucity of data from infrequently used keyboard keys.

This paper extends our earlier work in two ways. A stylometry biometric system has been developed to complement the keystroke one. And, for the first time, the data input to the system were obtained from students taking actual tests in a university course. Experiments on these data yielded keystroke, stylometry, and combined keystroke-stylometry performance results.

The organization of the paper is straightforward: section 2 describes the system procedures, section 3 the data collection process, section 4 the experimental performance results, and section 5 the conclusions and suggestions for future work.

2. Keystroke and Stylometry Systems

The keystroke system is described only briefly here since it has been described previously in detail [22]. The stylometry system is modeled after the keystroke system and the feature measurements are described below. The classification system is common to the keystroke and stylometry systems and is described in some detail. The combined keystroke-stylometry system simply concatenates the feature vectors from the two systems and uses the common classification system to obtain performance results.

An improved input system captures the keystroke timings and full input text in an XML file. The feature extractor parses each file creating both keystroke and stylometry feature vectors for later processing.

2.1. Keystroke System

The keystroke system consists of a raw keystroke data collector, a feature extractor, and a pattern classifier [22]. During a preprocessing phase, an outlier removal process eliminates key-press duration (dwell) and key transition (latency) times greater than two standard deviations from the mean over the whole dataset. This is particularly important for eliminating long transitions due to typing pauses from phone calls and other interruptions. To give each measurement roughly equal weight the features are standardized into the range 0-1 by clamping at plus and minus two standard deviations from the mean (previously, clamping was performed using the minimum and maximum values, and both methods yield comparable performance results). The 239 features, which are listed in [22], include means and standard deviations of the timings of key press durations and transitions, and percent useage of certain keys.

2.2. Stylometry System

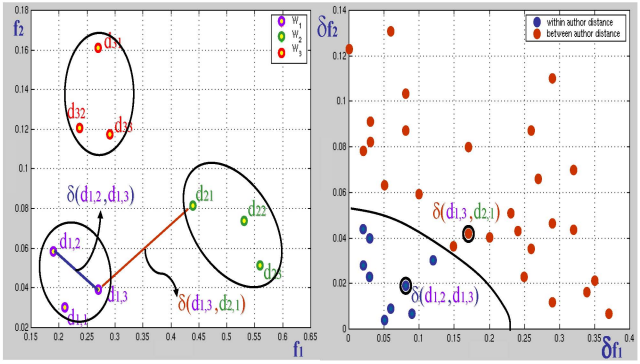
This system uses a set of linguistic features, basically a combination of the frequencies of the character-based features used in the keystroke biometric system [22], and the word and syntax level features used in an email stylometry study [27]. The features were normalized to be relatively independent of the text length – for example, *the number of different words (vocabulary) / total number of words* was used rather than simply *the number of different words*. The features were also chosen to show reasonable variation over a population of users – for example, some students will use a large vocabulary and others a small one. The 82 features – 49 character-based, 13 word-based, and 20 syntax-based features – are listed in the appendix. As in the keystroke system, the features are standardized into the range 0-1.

2.3. Combined Keystroke-Stylometry System

The combined system simply concatenates the 239 keystroke and the 82 stylometry features into a single vector of 321 features, and uses the common classification system to obtain authentication performance results.

2.4. Common Classification System

For authentication (verification), a vector-difference model transforms a multi-class problem into a two-class problem (Figure 1). The resulting two classes are "within-class (intra-person), you are authenticated" and "between-class (inter-person), you are not authenticated." This is a strong inferential statistics method found to be particularly effective for multidimensional feature-space problems [24].



(a) Feature space (b) Feature-difference space

Figure 1. Transformation from feature space (a) to feature distance space (b), adapted from [24].

To explain the dichotomy transformation process, take an example of three people $\{P_1,P_2,P_3\}$ where each person supplies three biometric samples. Figure 1 (a) plots the biometric sample data for these three people in two-dimensional feature space. This feature space is transformed into a feature-difference space by calculating vector distances between pairs of samples of the *same* person (*intra-person distances*, denoted by $x_{\oplus}$) and distances between pairs of samples of *different* people (*inter-person distances*, denoted by $x_{\varnothing}$). Let d_{ij} represent the individual feature vector of the i^{th} person's j^{th} biometric sample, then $x_{\oplus}$ and $x_{\varnothing}$ are calculated as follows:

$$x_{\oplus} = |d_{ij} - d_{ik}| \text{ where } i=1 \text{ to } n, \text{ and } j,k=1 \text{ to } m, j \neq k$$
$$x_{\varnothing} = |d_{ij} - d_{kl}| \text{ where } i,k=1 \text{ to } n, i \neq k \text{ and } j,l=1 \text{ to } m \quad (1)$$

where n is the number of people, m is the number of samples per person, and the absolute value is of the elements of these vectors. Figure 1 (b) shows the transformed feature distance space for the example problem.

If n people provide m biometric samples each, the numbers of intra-person and inter-person distance samples, respectively, are [24]:

$$n_{\oplus} = \frac{m \times (m-1) \times n}{2} \quad n_{\varnothing} = m \times m \times \frac{n \times (n-1)}{2} \quad (2)$$

In the authentication process, a user's keystroke sample requiring authentication is first converted into a feature vector. The difference between this feature vector and an earlier-obtained enrollment feature vector from this user is computed, and the resulting difference vector is classified as within-class (intra-person) for authentication or between-class (inter-person) for non-authentication. The *k*-nearest-neighbor method performs this classification by comparing this feature-difference vector against those in the training set.

To obtain system performance we simulate the authentication process of many true users trying to get authenticated and of many imposters trying to get authenticated as other users. This is done by using the numbers of the inter- and intra-person distances explained above. For example, if we have eight keystroke samples from each of 15 users, then (from the equation above) there are 420 intra-person distances to simulate true users and 6420 inter-person distances to simulate imposters. The feature distance space is populated similarly during training.

Receiver operating characteristic (ROC) curves are obtained by using a weighted procedure of the *k* nearest neighbors [26]. This procedure uses a linear rank weighting, assigning the first choice (nearest neighbor) a weight of *k*, second a weight of *k*-1, ... , and the k^{th} a weight of 1. The maximum score when all choices are within-class is $k+(k-1)+...+1 = k(k+1)/2$, and the minimum score is *0*. Now, consider that we authenticate a user if the weighted-within-class choices are greater or equal to *m*, where *m* varies from *0* to $k(k+1)/2$, and compute the (FRR, FAR) pairs for each *m* to obtain an ROC curve. The ROC curves in the experimental section below used ten nearest neighbors to provide weighted scores in the range 0-55 and thus 56 points on the ROC curve.

3. Data Collection

Data were collected from 40 students, predominantly juniors and seniors, in two sections of a spreadsheet modeling course in the business school of a four-year liberal arts college. The classes met in a 20-seat desktop computer laboratory where the exams were administered. Although this study investigated an online test-taking application, the data were captured in a classroom setting for greater experimental control.

The 40 students took four online short-answer tests of 10 questions each, and the tests took place at approximately two week intervals. The students were unaware that their data were being captured for experimental analysis. Data from students not completing all four tests or having problems with the input system were removed, resulting in complete data sets from 30 students, 17 male and 13 female, to be used in the experiments. What corresponds to failure to enroll problems were due to students having difficulty remembering their username and password or missing text input for a question because they clicked the "Next Question" button more than once. The data set, then, is comprised of only students that completed all four tests and all questions in each test.

The text lengths of the answers to a test ranged from 433 to 1831 words per test, with a mean of 966 and a median of 915 words. An average word length of five characters (six with spaces between words) yields roughly 6000 keystrokes per test as input to the keystroke system.

All the tests were taken on classroom Dell desktop computers with associated Dell keyboards. Training and testing on the same type of keyboard is optimal because it is known that keystroke data tends to vary for different keyboards, different environmental conditions, and different types of texts [7, 22].

4. Experimental Design and Results

Results were obtained on data from the 30 students who completed all four tests. Using data from different students for training and testing simulates an open system, while using data from all the students for both training and testing simulates a closed system.

Two open-system and two closed-system experiments (a total of four experiments) were performed on each of the keystroke, stylometry, and combined keystroke-stylometry systems. The experiment design is summarized in Table 1.

Table 1. Experimental design.

Experiment	Train and Test Samples	Numbers of Intra/inter Δ Samples
Experiment 1 Biometric Open-system	8 samples 5 answers combined 15 students	420/6420
Experiment 2 Biometric Open-system	4 samples 10 answers combined 15 students	90/1680
Experiment 3 Biometric Closed-system	4 samples 5 answers combined 30 students	180/6960
Experiment 4 Test Verification Closed-system	15 samples 10 answers combined 4 tests	420/1350

In the two open-system experiments, data from 15 students were used to train the system and data from the other 15 students were used to test the system and obtain performance results. Because the answers to the test questions could be short, several answers were combined to obtain the biometric samples. The numbers of intra- and inter-class samples is shown in the last column as computed from the formulas above for the difference-model classification scheme. An important advantage of this model is that a modest amount of data provides relatively large numbers of samples to evaluate system performance.

In the first experiment, five answers (half the test answers) were combined to obtain each sample, resulting in eight samples per student since each of the four tests contained ten questions for a total of 40 questions. With eight samples per student and 15 students for training and testing, there were 420 intra-person distances to simulate true users and 6420 inter-person distances to simulate imposters (equation 2).

In the second experiment, ten answers (all the answers of a test) were combined to obtain each sample, resulting in four samples per student. This yielded 90 intra-person distances to simulate true users and 1680 inter-person distances to simulate imposters (equation 2).

In the third experiment, the data samples from experiment 1 were used in a different way. The system was trained and tested on all the students, training on four samples and testing on the other four. With four samples per student and 30 students for training and testing, there were 180 intra-person distances to simulate true users and 6960 inter-person distances to simulate imposters (equation 2).

In the fourth experiment, the testing tried to verify the tests rather than the students. This was done primarily for stylometry to determine to what extent the students used words and other linguistic styles based on the test questions. With 15 samples per test and four tests for training and testing, there were 420 intra-person distances to simulate true users and 1350 inter-person distances to simulate imposters (equation 2).

4.1. Keystroke performance results

Figure 2 presents the keystroke-system ROC curves for experiments 1, 2, and 3. Performance improved in going from open-system experiment 1 to open-system experiment 2 which had longer data samples of twice as many keystrokes. Performance further improved in going from the open-system experiments to the closed-system experiment 3, even though experiment 3 had shorter data samples than experiment 2 (half the data) and half the number of training samples as experiment 1 (but the same as experiment 2). Figure 3 presents FRR and FAR versus *m*, the weighted *k*NN score, and clearly shows the experiment 3 closed-system EER at the crossover point. The EERs of the three keystroke experiments were 1.4%, 1.1%, and 0.55%, respectively.

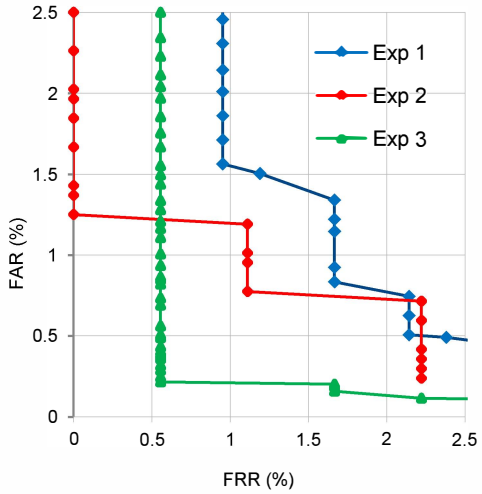


Figure 2. Keystroke ROC curves.

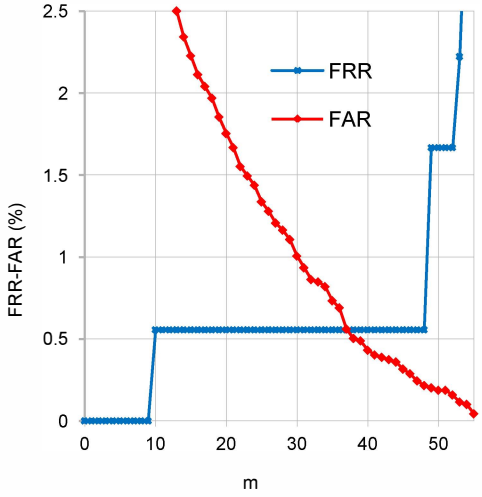


Figure 3. FRR/FAR versus *m* for Exp 3, EER=0.55%.

4.2. Stylometry performance results

Figure 4 presents the stylometry-system ROC curves for experiments 1, 2, and 3. Performance improved in going from open-system experiment 1 to open-system experiment 2 which had longer data samples of twice the text length. Performance deteriorated, however, in going from the open-system experiments to the closed-system experiment 3 which had shorter data samples than experiment 2 (half the data) and half the number of training samples as experiment 1 (but the same as experiment 2). The EERs of the three stylometry experiments were 40%, 33%, and 43%, respectively.

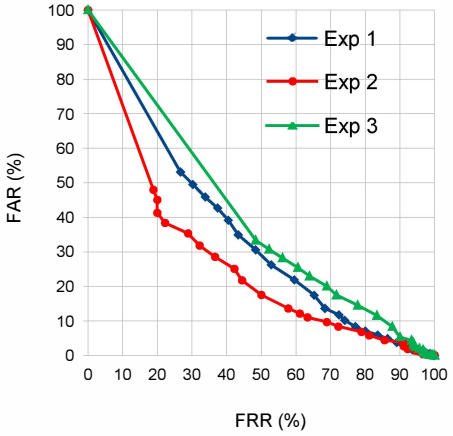


Figure 4. Stylometry ROC curves.

4.3. Combined keystroke-stylometry performance

The combined keystroke-stylometry system ROC curves are not presented here because they show a slight decrease in performance relative to that of the keystroke system alone, indicating the stylometry features do not provide the anticipated positive boost in performance, at least with the short text samples of this study.

4.4. Test verification performance results

The fourth experiment evaluated the capability of the keystroke and stylometry systems to verify the test rather than the student. It was hypothesized that there would be a correlation at the linguistic level between the students' answers to test questions and the test questions themselves, and that the stylometry system would recognize this correlation but the keystroke system would not. This hypothesis was based on the observation that many students repeated back portions of the questions in their answers, and the idea that they might give similar answers based on what they learned in class. Figure 5 presents the test-verification ROC curves for the keystroke and stylometry systems.

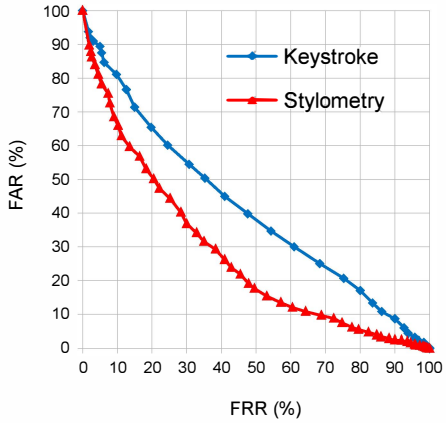


Figure 5. Test verification ROC curves.

As anticipated, the stylometry system discriminated among the tests and, in fact, did as well in discriminating among the tests as it did in discriminating among the students. The keystroke system did poorly on the test discrimination task. The EERs of the test-discrimination experiments were 43% for the keystroke system and 33% for stylometry.

4.5. Summary of performance results

A summary of the experimental performance results is shown in Table 2.

Table 2. Summary of experimental results.

Experiment	Train and Test Samples	Keystr EER	Stylo EER
Experiment 1 Biometric Open-System	8 samples 5 answers comb 15 students	1.4%	40%
Experiment 2 Biometric Open-System	4 samples 10 answers comb 15 students	1.1%	33%
Experiment 3 Biometric Closed-System	4 samples 5 answers comb 30 students	0.55%	43%
Experiment 4 Test Verification Closed-System	15 samples 10 answers comb 4 tests	43%	33%

The first three experiments measured the performance of the keystroke and stylometry biometric systems. For the keystroke system, performance increased from experiment 1 to experiment 2 with the doubling of the data size, and further increased in going to the closed-system experiment 3 even with the decrease in data size back to that of experiment 1 and an increase in population size from 15 to 30 students, indicating that going to a closed system is more important than the increase in data size. For the stylometry system, performance increased from experiment 1 to 2 with the doubling of the data size, but then decreased in going to the closed-system experiment 3 with the data size reverting back to that of experiment 1, indicating that going to a

closed system of more students is less important than the increase in data size for stylometry.

The fourth experiment measured the performance of verifying the test and was not a biometric experiment. The stylometry system could verify the tests as well as it could verify the students, but the keystroke system had minimal test discrimination capability.

5. Conclusions

The results obtained on the keystroke system are an improvement over those previously reported and, for the first time, the data were obtained from students taking actual tests. The best performance reported here is 0.55% EER on a closed system of 30 students, while previously reported performance was 1.0% EER on a closed system of 14 students [26]. We have also shown the degree to which performance increases as the size of the data samples (number of keystrokes) increases, and how performance in a closed system is superior to that in an open system.

The performance of the keystroke biometric system is far superior to that of the stylometry one. While the keystroke and stylometry biometrics are both behavioral biometrics, they operate at different cognitive levels. The keystroke biometric operates at essentially an automatic motor control level. Stylometry, however, operates at a higher cognitive level, and because it primarily involves word and syntax-level units, much longer text passages are required relative to those required by the keystroke biometric.

To obtain system performance in this study we simulated the authentication process of many true users trying to get authenticated and of many imposters trying to get authenticated as other users. An important advantage of this vector-difference model is that it provides relatively large numbers of inter- and intra-person distance samples. Although test-taker authentication in real time would not be possible with the described technique due to the significant amount of input required (half or full test), delayed authentication with batch processing should be sufficient for university and HEOA requirements.

Important parameters in authorship attribution methods are the length and number of training and testing texts, and the number of potential authors [20]. Another important factor discovered in this stylometry study was the relationship between the texts under study and how the texts are produced. For example, we found a relatively strong correlation between the test answers and the test questions producing the answers. Therefore, better performance results would likely be obtained from student essays on a variety of topics, as might be obtained from students in an English class, although two students who happen to choose the same or similar topic may present a problem. Another authorship study parameter may be the medium, where perhaps the idiosyncratic styles of the users of the newer mediums like email and tweets have allowed stylometry studies to be somewhat successful [11, 16, 27].

Future work on improving stylometry in test taking applications might investigate the use of idiosyncratic features like the fraction of misspelled words. The use of longer text passages and those on different topics, such as essays in English classes, might also be explored, as well as different ways of fusing the keystroke and stylometry results. Finally, while the experiments reported here used actual test data, the authentication process itself was simulated by enumerating all the combinations of sample pairs, so future work might explore an actual authentication process in a student testing environment.

References

[1] AdmitOne Security Inc. http://www.admitonesecurity.com/, accessed May 2011.

[2] S.S. Bender and H.J. Postley. Key sequence rhythm recognition system and method. U.S. Patent 7,206,938, 2007.

[3] F. Bergadano, D. Gunetti, and C. Picardi. User authentication through keystroke dynamics. ACM Trans. Info. & System Security, 5(4):367-397, 2002.

[4] BioChec. http://www.biochec.com/, accessed May 2010.

[5] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior. *Guide to biometrics*. NY: Springer, 2004.

[6] R. Giot, M. El-Abed, and C. Rosenberger. Keystroke dynamics with low constraints svm based passphrase enrollment. IEEE Int. Conf. Biometrics: Theory, Applications, and Systems (BTAS), 2009.

[7] D. Gunetti and C. Picardi. Keystroke analysis of free text. ACM Trans. Info. and System Security, 8(3):312-347, 2005.

[8] Higher Education Opportunity Act (HEOA) of 2008, http://www2.ed.gov/policy/highered/leg/hea08/index.html, accessed May 2011.

[9] L. Jin, X Ke, R. Manuel, and M. Wilkerson. Keystroke dynamics: a software based biometric solution. 13th USENIX Security Symposium, 2004.

[10] M.L. Jockers and D.M. Witten. A comparative study of machine learning methods for authorship attribution. Literary and Linguistic Computing, 25(2): 215-223, 2010.

[11] R. Layton, P. Watters, and R. Dazeley. Authorship attribution for twitter in 140 characters or less. Second Cybercrime and Trustworthy Comp. Workshop, 1-8, 2010

[12] J. Leggett and G. Williams. Verifying identity via keystroke characteristics. Int J. Man-Machine Stud., 28(1):67-76, 1988.

[13] J. Leggett, G. Williams, M. Usnick, and M. Longnecker. Dynamic identity verification via keystroke characteristics. Int. J. Man Machine Studies, 35(6): 859-870, 1991.

[14] F. Montrose, M.K. Reiter, and S. Wetzel. Password hardening based on keystroke dynamics. Int. J. Info. Security, 1(2): 69-83, 2002.

[15] M.S. Obaidat and B. Sadoun. Keystroke dynamics based authentication. In *Biometrics: Personal Identification in Networked Society* by A.K. Jain, R. Bolle, and S. Pankanti. New York: Springer, 213-230, 1999.

[16] D. Pavelec, L.S. Oliveira, E. Justino, F.D. Nobre Neto, and L.V. Batista. Compression and stylometry for author

identification. Int. Joint Conf. Neural Networks, 2445-2450, 2009.
[17] A. Peacock, X. Ke, and M. Wilkerson. Typing patterns: a key to user identification. IEEE Security & Privacy, 2(5): 40-47, 2004.
[18] K. Revett. Chap 4: Keystroke dynamics, 73-136. *Behavioral biometrics: a remote access approach*, Wiley, 2008.
[19] R. N. Rodrigues, G.F.G. Yared, C.R. Costa, J.B.T. Yabu-Uti, F. Violaro, and L.L. Ling. Biometric access control through numerical keyboards based on keystroke dynamics. Lecture Notes in Computer Science, 3832: 640-646, 2006.
[20] E. Stamatatos. A survey of modern authorship attribution methods. J. Am. Soc. Info. Science and Tech., 60(3): 538–556, 2009
[21] Stylometry. Wikipedia. http://en.wikipedia.org/wiki/Stylometry, accessed May 2011.
[22] C.C. Tappert, S. Cha, M. Villani, and R.S. Zack. A keystroke biometric system for long-text input. Int. J. Info. Security and Privacy (IJISP), 4(1): 32-60, 2010.
[23] M. Villani, C. Tappert, G. Ngo, J. Simone, H. St. Fort, H-S Cha. Keystroke biometric recognition studies on long-text input under ideal and application-oriented conditions. Computer Vision & Pattern Recognition Workshop on Biometrics, New York, 2006.
[24] S. Yoon, S-S Choi, S-H Cha, Y. Lee, and C.C. Tappert. On the individuality of the iris biometric. Int. J. Graphics, Vision & Image Processing, 5(5): 63-70, 2005.
[25] E. Yu and S. Cho. Keystroke dynamics identity verification – Its problems and practical solutions. Computers & Security, 23(5): 428-440, 2004.
[26] R.S. Zack, C.C. Tappert and S.-H. Cha. Performance of a long-text-input keystroke biometric authentication system using an improved k-nearest-neighbor classification method. IEEE 4th Int. Conf. Biometrics: Theory, Applications, and Systems (BTAS), 2010.
[27] R. Zheng, J. Li, H. Chen, and Z. Huang. A framework for authorship identification of online messages: writing-style features and classification techniques. J. Am. Soc. Info. Science and Tech., Feb 2006.

Appendix

Table A1. List of stylometry features.

Character-based features:
1. number of alphabetic characters/total number of characters
2. number of uppercase alphabetic characters/ number of alphabetic char
3. number of digit characters/total number of characters
4. number of space characters/total number of characters
5. number of vowel (a,e,i,o,u) characters/number of alphabetic characters
6. number of "a" (upper or lowercase) characters/number of vowel characters
7. number of "e" characters/number of vowel characters
8. number of "i" characters/number of vowel characters
9. number of "o" characters/number of vowel characters
10. number of "u" characters/number of vowel characters
11. number of most frequent consonants (t,n,s,r,h)/number of alph char
12. number of "t" characters/number of (t,n,s,r,h)
13. number of "n" characters/number of (t,n,s,r,h)
14. number of "s" characters/number of (t,n,s,r,h)
15. number of "r" characters/number of (t,n,s,r,h)
16. number of "h" characters/number of (t,n,s,r,h)
17. number 2nd most frequent consonants (l,d,c,p,f)/number of alph char
18. number of "l" characters/number of (l,d,c,p,f)
19. number of "d" characters/number of (l,d,c,p,f)
20. number of "c" characters/number of (l,d,c,p,f)
21. number of "p" characters/number of (l,d,c,p,f)
22. number of "f" characters/number of (l,d,c,p,f)
23. number 3rd most frequent consonants (m,w,y,b,g)/number of alph char
24. number of "m" characters/number of (m,w,y,b,g)
25. number of "w" characters/number of (m,w,y,b,g)
26. number of "y" characters/number of (m,w,y,b,g)
27. number of "b" characters/number of (m,w,y,b,g)
28. number of "g" characters/number of (m,w,y,b,g)
29. number of least frequent consonants (j,k,q,v,x,z) / number of alph char
30. number of consonant-consonant digrams/total number alph digrams
31. number of "th" digrams/number consonant-consonant digrams
32. number of "st" digrams/number consonant-consonant digrams
33.number of "nd" digrams/number consonant-consonant digrams
34. number of vowel-consonant digrams/total number alph digrams
35. number of "an" digrams/ number of vowel-consonant digrams
36. number of "in" digrams/ number of vowel-consonant digrams
37. number of "er" digrams/ number of vowel-consonant digrams
38. number of "es" digrams/ number of vowel-consonant digrams
39. number of "on" digrams/ number of vowel-consonant digrams
40. number of "at" digrams/ number of vowel-consonant digrams
41. number of "en" digrams/ number of vowel-consonant digrams
42. number of "or" digrams/ number of vowel-consonant digrams
43. number of consonant-vowel digrams/total number of alphabet digrams
44. number of "he" digrams/ number of consonant-vowel digrams
45. number of "re" digrams/ number of consonant-vowel digrams
46. number of "ti" digrams/ number of consonant-vowel digrams
47. number of vowel-vowel digrams/total number of alphabet letter digrams
48. number of "ea" digrams/total number of vowel-vowel digrams
49. number of double-letter digrams/total number of alphabet letter digrams

Word-based features:
1. number of one-letter words/total number of words
2. number of two-letter words/total number of words
3. number of three-letter words/total number of words
4. number of four-letter words/total number of words
5. number of five-letter words/total number of words
6. number of six-letter words/total number of words
7. number of seven-letter words/total number of words
8. number of long words (eight or more letters)/ number of words
9. number of short words (one to three letters)/ number of words
10. average word length = number letters in all words/total number of words
11. number of different words (vocabulary)/total number of words
12. number of words occurring once/total number of words
13. number of words occurring twice/total number of words

Syntax-based features:
1. number of the eight punctuation symbols (.,?!;:'")/ total number of char
2. number of periods (.)/total number of the eight punctuation symbols
3. number of commas (,)/total number of the eight punctuation symbols
4. number of "?" and "!"/total number of the eight punctuation symbols
5. number of semicolons (;) and colons (:)/total number punctuation symbols
6. number of single (') and double quotes (")/number punctuation symbols
7. total number of non-alphabetic, non-punctuation, and non-space characters (0,1,2,3,4,5,6,7,8,9,@,#,$,%,etc.)/total number of characters
8. number of digit char/number of non-alph, non-punct, and non-space char
9. total number of articles (a, an, the)/total number of words
10. total number of "the" articles/total number of articles
11. total number of "a" or "an" articles/total number of articles
12. total number of common conjunctions/total number of words
13. total number of common interrogatives/total number of words
14. total number of common prepositions/total number of words
15. number of first-person personal pronouns/number of personal pronouns
16. number of second-person personal pronouns/number personal pronouns
17. number of third-person personal pronouns/ number of personal pronouns
18. total number of personal pronouns/total number of words
19. average number of characters per sentence
20. average number of words per sentence