When linguistics meets computer science: Stylometry

Jean Langlois-Berthelot

Volume 5 Issue 2, 2021, pp. 51-61 doi: 10.22363/2521-442X-2021-5-2-51-61

Original Research

When linguistics meets computer science: Stylometry and professional discourse

by Jean Langlois

Jean Langlois Sciences Po, France jean.langlois@sciencespo.fr
Article history Received February 28, 2021 | Revised June 1, 2021 | Accepted June 14, 2021
Conflicts of interest The authors declared no conflicts of interest
Research funding No funding was reported for this research
doi 10.22363/2521-442X-2021-5-2-51-61
For citation Langlois, J. (2021). When linguistics meets computer science: Stylometry and professional discourse. *Training, Language and Culture, 5*(2), 51-61.

In an 1887 article, The Characteristic Curves of Composition, published in the journal 'Science' and resulting from discussions with the logician Augustus de Morgan, Mendenhall (1888) asserted that the length of words was a characteristic capable of distinguishing authors of a literary text. This study is often considered one of the founders of stylometry as a discipline. This paper analyses the development of stylometry and its use as a computerised analytical tool and explores its potential as a way of identifying authors of online professional communication by the vocabulary and style they use. The objective of this paper is to explain the concept of stylometry as an academic methodology, how it has been adapted to computers and how it is used in online investigation of author and narrative identity. The methodology is based on secondary research to explain how stylometry can be used in author definition and attribution and identification of texts. It goes on to analyse the types of stylometric entities and examines the role of computers in stylometry and its application to professional discourse. The study concludes that although the use of computers is an important quantitative tool in stylometric research, in the end it is human judgement that counts.

KEYWORDS: *stylometry, professional discourse, lexis, structural analysis, content analysis, Enron*

1. INTRODUCTION

The moment of crystallisation and theorisation of the methods of stylometry came in the second half of the 19th century with academics close to the *Cambridge and Dublin Mathematical Journal*. The researchers who participated in the early development of stylometry were generally both philosophers and mathematicians and participated in the mathematical-logical foundation of computer science. This paper traces the history of stylometry and its current uses and explores its potential application to professional discourse.

Stylometry is based on the principle that each author or organisation develops its own linguistic style and expression. Therefore, documents from an individual or an organisation can be analysed to identify most probable authorship. Stylometry is also a valuable tool in professional discourse, not just for identifying who wrote what and when and why, but also for identifying lexical and grammati-

Volume 5 Issue 2, 2021, pp. 51-61 doi: 10.22363/2521-442X-2021-5-2-51-61

'Stylometry is also a valuable tool in professional discourse, not just for identifying who wrote what and when and why, but also for identifying lexical and grammatical devices used in written communication, which can be compared with what is taught to students in business language textbooks'

cal devices used in written communication, which can be compared with what is taught to students in business language textbooks. In this sense, it is potentially an important source of authentic language used in professional discourse.

This study examines the background of stylometry and analyses the key types of stylometric entities used to identify and verify manuscript authorship. It then goes on to discuss the advantages and disadvantages of automatic word processing and concludes with its application to professional and organisational discourse. Based on secondary research, the study also includes a case study of Enron, the leading US finance firm that crashed in 2001, the responsibility for which was identified through the stylometric analysis of employees' emails using the Adaboost and SVM classifiers. It also examines the difference between the professional discourse language exponents taught in business language classes and the use of authentic language in business meetings to ultimately draw conclusions about the function of each in teaching professional discourse. Stylometric analysis is potentially an important technology not just in verifying literary narrative, but also revealing identity of key actors in organisational fraud and mismanagement and identifying authentic business language which can serve as a teaching resource.

2. THEORETICAL BACKGROUND

As Sigelman and Jacoby (1996) point out, stylometry is a very reliable technique and is widely used in literary studies but not yet in business communication. Stylometry is based on a principle widely confirmed by scientific literature: the linguistic style of an individual is so particular that it can be used to distinguish their writing. Even a writer who flaunts an abstruse vocabulary will also need to use many mundane words. Stylisticians regard style as a general predisposition toward a particular mode of expression rather than an invariant habit or constant. Stylometry is based on a very simple theory: just as each person has their own fingerprints or their own genome, each individual has their own style of writing and expression. The most obvious examples of this kind of style specificity are repetitions of the same spelling mistakes.

Lowe and Matthews (1995) call stylometry 'stylistic statistics' and define it as the application of mathematical methods to extract quantitative measures from a text. The main data that stylometry focuses on is the word and its articulation in a sentence (hence the issue of punctuation). Word and punctuation are the raw materials of this science.

Holmes (1994), one of the most influential researchers in the field of stylometry, explains that each text can be defined by a set of measurable statistical characteristics: if several documents emanating from the same individual have the same characteristics, then we can think that the author is making recurring use of a particular style. Holmes (1994) explains that there are no better criteria for making a comparison work between authors than those used in stylometric analysis. *'The lexical level is the obvious place to initiate stylistic investigations, since questions about style are essentially comparative and more data exist at the lexical level than at any other in the form of computed concordances'* (Holmes, 1994, p. 87-88).

In order to be able to study the stylistic characteristics of a text by computer, they must meet a number of conditions described by Bailey (1969, p. 219): they should be *'salient, structural, frequent and easily quantifiable, and relatively immune from conscious control'*. When we measure such characteristics, we try to highlight the uniqueness of an individual's handwriting and distinguish what makes their style unique from that of another individual. As Holmes (2004) reminds us, stylometry makes it possible to distinguish be-

tween true stylistic differences and variations that are only due to chance. The best tool to perform stylometric analysis of a text, that is to say concretely to derive interpretable statistical indicators, is computer science. However, computational linguistics software cannot analyse a text with the same depth as a human researcher.

The objects of computer analysis are of two types: the so-called 'raw' data (punctuation, character strings, syllables, etc.) and data which is in the order of content (themes, lexical fields, formulations, language registers, vocabulary, etc.).

Writing style is an unconscious habit, which is different from author to author in that to express an idea an individual will have personal use of grammar, words and punctuation. Although the style of writing can change a bit over time, each author actually has a recognisable stylistic tendency.

Almost all of the literature on stylometry shows that this methodology has three main objectives.

1. The question of authorship: who is the author of this text?

2. The question of verifiability: is this text from author X (specialists call it 'matching'), is it from author Y or from X and Y?

3. The question of characterisation also called 'profiling': what constitutes X's style?

To these we can add a fourth and fifth objective.

4. How can stylometry be applied to professional discourse in a way that can be used to identify authors of social media posts and emails?

5. How can stylometry be used in written professional discourse to reinforce accepted ways of expressing content in emails and social media posts?

The Internet is supposed to be a space of great freedom, which is particularly linked to the supposedly anonymous nature of our actions on the web. However, not only do we know that our navigation path can be easily traced thanks to our IP address in particular, but the writings that we publish in various social networks, forums, commercial sites, etc. constitute different traces of our passage in these virtual places and above all can help to reconstitute our style of writing.

'Writing style is an unconscious habit, which is different from author to author in that to express an idea an individual will have personal use of grammar, words and punctuation. Although the style of writing can change a bit over time, each author actually has a recognisable stylistic tendency'

3. DEFINING THE AUTHOR

3.1. Author verification

Paternity verification using stylometry for online documents (e.g. emails, tweets) poses significant challenges due to the unstructured nature of these documents. In addition, a major requirement is that (repeated) authentication decisions must occur over a short period of time or over short texts or messages. Stylometric analysis of short messages is difficult due to the limited amount of information available for decision making. Likewise, most of the stylometry analysis approaches proposed in the literature use a relatively large document size which is not acceptable for continuous authentication.

Another important challenge to address when using stylometry is the threat of counterfeiting. An adversary with access to a user's sample writing may be able to efficiently reproduce many existing stylometric features. It is essential to integrate specific mechanisms into the authentication system that would mitigate tampering attacks.

Author verification follows a typical biometric verification process, in which the identity of an author is verified by one-to-one matching. Some researchers have studied authorship verification as a similarity detection problem, where the problem is to determine the degree of similarity given by two texts, by measuring the distance between them. Other researchers have studied this question as a problem with one or two classes, one class consisting of documents written by the author and a second class consisting of documents written by other authors.

Volume 5 Issue 2, 2021, pp. 51-61 doi: 10.22363/2521-442X-2021-5-2-51-61

3.2. Attribution and identification

Author attribution follows a typical biometric identification process, where the system recognises an author by a 'one to several' comparison. The process consists of extracting features from sample texts and labelling the classes according to the authors of the documents. Typical entity categories include lexical, semantic, syntactic, and application-specific characteristics. This can also apply to SFL and professional discourse where different types and styles of online communication can be attributed to authors and organisations by the type of language they use. This is particularly important in the analysis of text classification.

Author attribution is similar to text classification. A key difference, however, is that author attribution is subject independent, while in text classification class labels are based on the subject of the document and features include subject dependent words. By being able to recognise the use of subject dependent words relevant to a particular type of communication, stylometry can be used to improve knowledge and application of appropriate means of expression in emails and social media posts.

Despite the significant progress made in identifying an author among a small group of people, it remains difficult to identify an author when the number of applicants increases or when the sample text is short as in the case of emails or online messages. For example, while Chaski (2005) reported an accuracy of 95.70% in his work on author identification, the evaluation sample came from a corpus of only 10 authors.

Likewise, Hadjij et al. (2009) obtained, using the K-means algorithm for author identification, a classification accuracy of 90% with only 3 authors; the rate decreased to 80% when the number of authors increased to 10. Hadjij et al. (2009) also proposed another approach called *Author-Miner*, which consists of an algorithm that captures frequent lexical, syntactic, structural and content-specific patterns. The experimental evaluation used a subset of the Enron dataset, varying from 6 to 10 authors, with 10 to 20 text samples per author. The accuracy of perpetrator identification decreased from 80.5% to 77% as the size of the perpetrator population increased from 6 to 10. Hadjij et al. (2009) used the C4.5 and SVM classifiers to determine authorship and assessed the proposed approach using a subset of three authors from the Enron dataset. They obtained correct classification rates of 77% / 71% for sender identification, 73% / 69% for sender-recipient identification, and 83% / 83% for sender identification.

'Author attribution is similar to text classifica-tion. A key difference, however, is that author attribution is subject independent, while in text classification class labels are based on the subject of the document and features include subject dependent words'

The issue of author attribution is particularly important in identifying the actual author of a disputed anonymous document. In the literature, author identification is considered to be a problem of text categorisation or text classification. The process begins with data cleansing followed by feature extraction and normalisation. Each suspect document is converted into a feature vector; the suspect represents the class label. The extracted entities are classified into two groups: training and testing sets. The training set is used to develop a classification model while the test set is used to validate the developed model assuming that the class labels are not known. Common classifiers include decision trees, neural networks and SVM (Support Vector Machine). Author attribution studies differ in terms of the stylometric characteristics used and the type of classifiers employed. Cho et al. (2013) propose two approaches that attempt to extract authorship from emails in the context of computer forensics. The authors extract various characteristics from email documents, including linguistic characteristics, header characteristics, linguistic models and structural characteristics. All of these features are used with the SVM learning algorithm to assign authorship of email messages to an author.

In their famous study Chen et al. (2011) develop a framework for the identification of the author in online messages to address the problem of identity tracing. Within this framework, four types of writing style features (lexical, syntactic, structural, and content-specific) are extracted from English and Chinese online newsgroup posts. A comparison was made between three classification techniques: decision tree, SVM and backpropagation neural networks. Experimental results showed that this framework is able to identify authors with satisfactory accuracy of 70-95% and that the SVM classifier outperformed the other two.

3.3. Author characterisation

Author characterisation is used to detect sociolinguistic attributes such as gender, age, occupation and educational level of the potential author of an anonymous document. Stylistics or the study of stylometric characteristics shows that individuals can be identified by the redundancy of their writing styles. Olsen (1993) attempted to define what constitutes the 'singular style' of a given person: an individual's writing style is defined by the terms used, the selection of special characters and the composition of sentences. Studies in the literature show that there are no such optimised functionalities applicable to all fields.

4. FOUR TYPES OF STYLOMETRIC ENTITIES

4.1. Lexical features

A text can be seen as a sequence of 'tokens' which are grouped into sentences. A token can be a word, a number or a punctuation mark. Many studies of author attribution rely primarily on simple measures such as the number and length of sentences and the number and length of words. The advantage of these features is that they can be applied to any corpus in any language and without additional requirements (Schuster et al., 2020).

Lexical features allow you to understand how an individual or a business organisation or a particular profession uses characters and words. For example, these characteristics may be represented in the frequency of special characters, the total number of capital letters used, the average number of characters per word, the average number of characters per sentence and compliance with punctuation rules.

'Another method to define a set of lexical features consists in extracting the most frequent words in the corpus. It is also possible to provide various clerical error metrics to capture idiosyncrasies of an author's style. To do this, we must define a set of spelling errors and formatting and propose a methodology to automatically extract these measures using a spell checker'

A text can be thought of as a sequence of characters. Different character-level metrics can then be defined, including number of alphabetic characters, total number of characters, number of upper- and lower-case characters, letter frequency and number of punctuation marks.

Vocabulary richness functions quantify the vocabulary diversity of a text. Some examples of this measure are: the V/N ratio (V is the size of the vocabulary and N is the total number of tokens in the text), the measure of Yule, the number of hapax legomena (words appearing once) and the number of hapax dislegomena (words appearing twice). However, the richness of the vocabulary strongly depends on the length of the text.

Various functions have been proposed to achieve stability over text length, including Yule measure, Simpson measure, Sichel measure, Brunet measure, and Honore measure (Eder, 2017).

Another method to define a set of lexical features consists in extracting the most frequent words in the corpus. It is also possible to provide various clerical error metrics to capture idiosyncrasies of an author's style. To do this, we must define a set of spelling errors (omissions and insertions of letters) and formatting (capital letters) and propose a methodology to automatically extract these measures using a spell checker.

Volume 5 Issue 2, 2021, pp. 51-61
doi: 10.22363/2521-442X-2021-5-2-51-61

4.2. Syntactic characteristics

Holmes (1998) defines syntactic characteristics as the patterns used to form sentences. This category of entities includes the tools used to structure sentences. These include punctuation and function words. Function words are common words (articles, prepositions, pronouns) like *while, upon, though, where, your.*

4.3. Structural features

Structural characteristics are useful in understanding how an individual organises the structure of their written speech. For example, how sentences are organised in paragraphs and paragraphs in a given document. Structural characteristics are generally suggested for email author attribution. In addition to general structural characteristics, several researchers have used specific characteristics of emails, such as the presence/absence of greetings and signatures and their position in the body of the email (Savoy, 2020).

4.4. Content-specific features

Content-specific features can be used to characterise the authors of certain texts in interaction situations such as in discussion forums using keywords. Work on the characterisation of authors has targeted the determination of various traits or characteristics of an author such as sex, age or level of education according to the use of a particular keyword repeatedly or of a specific formulation. The general approach is to create sociolinguistic groups from documents written by the same population, then to deduce to which group(s) an anonymous document could be linked.

Cheng et al. (2011) studied author gender identification from text using the Adaboost and SVM classifiers to analyse 101 lexical features, 10 syntactic features, 13 structural features, and 392 functional features.

In December 2001 a leading US financial firm, Enron, collapsed amidst accusations of fraud. In order to trace what had caused the crash and who was involved, the FERC subpoenaed for examination of 600,000 emails generated by 158 employees. The evaluation of the proposed approach involving 108 authors from the Enron dataset yielded classification accuracies of 73% and 82.23%, respectively, for Adaboost and SVM.

'As Fortier (1995) recalls, software is capable of analysing large amounts of information in record time, the tasks performed by specialised software are long and laborious, even impossible for humans alone. Moreover, unlike human analysis, the automated processing of data by specialised software is objective a priori. The algorithms allow the examination of a whole text without special attention being drawn to a particular passage more than on another'

Abbasi and Chen (2008) analysed the individual characteristics of participants in an extremist group web forum using the decision tree and SVM classifiers. The experimental evaluation gave success rates of 90.1% and 97% in identifying the correct author among 5 possible individuals for the decision tree and the SVM, respectively.

Kucukyilmas et al. (2008) used the k-NN classifier to identify a user's gender, age, and educational background. The experimental evaluation of 100 participants grouped by sex (2 groups), age (4 groups) and educational background (10 groups) gave precision of 82.2%, 75.4% and 68.8%, respectively.

As Fortier (1995) recalls, software is capable of analysing large amounts of information in record time, the tasks performed by specialised software are long and laborious, even impossible for humans alone. Moreover, unlike human analysis, the automated processing of data by specialised software is objective a priori. The algorithms allow the examination of a whole text without special attention being drawn to a particular passage more than on another. *'An algorithm has the immense advantage of not being subject to distractions, nor to preconceived ideas'* (Fortier, 1995, p. 101).

5. DISCUSSION

5.1. Advantages of automatic word processing

There are indeed many advantages of automatic word processing. Burrows and Craig (1994) have shown in a long article on scientific debates in the field of textual analysis, that statistical computer analysis of literary texts has been an unprecedented revolution, not so much in that it brings a technical renewal to the field of study, but quite simply in that it has given a firmer basis to many methodological debates which, until then, were lost in conjecture (Vartanova et al., 2020; Zvereva, 2021).

Olsen (1993) reminds us that the automatic analysis of texts must be defined in a simple and limited way. It does this by highlighting the aspects of a text or of a series of texts which would be difficult to see with the naked eye. The computer is therefore above all what Olsen (1993) describes as an accelerator and a facilitator.

Automatic data processing can be used to analyse data from entire literary works rather than from individual texts themselves. The place of humans is then much more important, their involvement is more important, they must indeed collect, organise, classify and process data, etc. prior to automated processing.

In their research on the importance of female characters in French-speaking African literature, Ormerod et al. (1994) used several computer software programmes. The three researchers have assembled a representative corpus of ten novels written by male authors and ten written by women. The data submitted for the software review consisted of an exhaustive list of the characters in these twenty novels.

The characters were assigned three ratings from 1 to 5 (according to pre-established criteria): one corresponding to the importance given to the character, another to the power conferred on them in the social and professional field, and the last to their attitude in the novel. It is therefore at this level that the most important part of human intervention was located before automated processing. The researchers then highlighted the difference between texts by male authors and those by female authors.

Laffal (1995) has worked extensively on analysing the works of Irish writer Jonathan Swift. As this author not only wrote in English, Laffal (1995) also translated words from works that were in languages other than English and replaced proper names with names or places, whichever they referred to. The scholar identified the various words whose spelling had changed since the writing of these texts and associated them with corresponding words according to modern spelling. As Laffal (1995) explains at length in his article, he even had to intervene during the analysis to counter the problems of polysemy. From then on, he had to use two software: one read the text to be analysed and marked all words which had more than one meaning in the dictionary, and the other advanced through the marked text, stopping at each marked word with a display of numbered dictionary choices. The human editor selected the proper meaning by keying the relevant number.

5.2. Criticisms and limits

For some computer text analysis experts, led by Olsen (1993), computing has a lot to offer to stylometry, but software is often misused and the results do not have the impact they might have. The scholar considers that it is necessary to reassess the role of informatics in the analysis of the literature and to move in new directions. He quotes Potter (1989) who says that specialists using computer science in literature too often tend to make their reports very 'technical', which does not help to gain a readership of literary works.

There would then be a 'complex encountered by computer engineers' seeking to hide behind a technical vocabulary – too often poorly mastered. Potter (1989) also notes that this type of study is mostly limited to a small number of works.

Brunet (2003), for his part, raises the dangers of what he calls 'statistical stubbornness'. When a researcher wishes, for example, to determine the authorship of a text, he will begin by formulating a hypothesis first and then subjecting the text to tests in order to verify or refute his hypothesis. When it does not achieve the desired results and does not want its efforts to be in vain, there is a high risk

Volume 5 Issue 2, 2021, pp. 51-61 doi: 10.22363/2521-442X-2021-5-2-51-61

'To date no known algorithm makes it possible to grasp whether a given word is used in a figurative or literary sense. To do this, it would in fact be necessary for the researcher to carry out encoding work beforehand, which constitutes a long and tedious task. According to Miall (1995), to predict a new era in which a computer would be able to grasp and understand a literary work would underestimate the complexity of the process of reading a text by a human being'

that it will push hard and try to interpret the results in a way that makes them say the right thing. Brunet (2003) recalls that there is a strong tendency in stylometry to give figures an almost divine superiority over words because they seem absolute, *'but this apparent incontrovertibility, however impressive, often conceals relative and contingent procedures that have nothing essential about them'* (Brunet, 2003, p. 70).

According to Olsen (1993), the main errors that are made by people using automated processing in the field of stylometry are not technical, but rather theoretical and methodological. Olsen (1993) explains that computer text analyses are usually done on the basis of too simple things like word length and what he describes as 'type-token' ratios, when these measurements give unsatisfactory results on their own. This is also the opinion of Miall (1995), who believes that *'the frequencies of words, collocations, or particular stylistic features tell us rather little about the literary qualities of a text, since these aspects of a text find their meaning only within the larger and constantly shifting context constituted by the reading process'* (Miall, 1995, p. 275).

As Fortier (1995) explains, it is by no means easy to transform textual qualities into analysable statistics. Although the texts are composed of words, their effects are produced by phenomena of a higher and more complex order. To date no known algorithm makes it possible to grasp whether a given word is used in a figurative or literary sense. To do this, it would in fact be necessary for the researcher to carry out encoding work beforehand, which constitutes a long and tedious task. According to Miall (1995), to predict a new era in which a computer would be able to grasp and understand a literary work would underestimate the complexity of the process of reading a text by a human being.

Nonetheless, one can agree with Olsen (1993) that automated word processing allows data to be updated which can form the basis of human deeper work. It would seem that the approach of using computers to analyse the linguistic and symbolic environment – the collective and social elements of language – in order to understand individual texts and rhetorical stances, suggests that computer analysis of text should play a central and well-defined role in our understanding of text. On the other hand, while some aspects of literary texts are quantifiable, others can never be.

5.3. Application to professional discourse

As mentioned above, computer analysis of text can add to professional discourse in two ways – security and modelling. As we saw in the analysis of the Enron email database, computer analysis can help identify who wrote what in emails and potentially in social media, thereby enhancing openness and transparency of authorship. Of more value to teachers and researchers, however, is the potential of computer analysis to identify areas of content and lexis used by native speakers to convey information and understanding through emails and other social media. In doing so, learners will improve both their ability and fluency in written communication through electronic media. Another factor in professional discourse is that the use of language and particularly jargon changes rapidly and can be difficult to keep up with, even by native speakers. The value of computer analysis is to identify these changes and, with professional human support, explain them and teach them in a way that learners will not only gain new informa-

'Computer analysis of text can add to professional discourse in two ways – security and modelling. As we saw in the analysis of the Enron email database, computer analysis can help identify who wrote what in emails and potentially in social media, thereby enhancing openness and transparency of authorship'

tion but know what language it is appropriate to use, and in what type and style of communication. An important point to consider is the difference between more formal and informal communication. Social media communication tends towards the informal whereas email can be both formal or informal. Understanding and observing the difference in language usage when reading or writing is important as it has an influence on cultural understanding and on building good relations.

One of the key areas of research in language and professional discourse is the relationship between what is taught in textbooks and what takes place in real life meetings, described as 'authentic' English. Authenticity is defined as something of undisputed origin and not a copy, based on facts, accurate and reliable. Williams (1988) identified the language used in real business meetings she attended: *'On reading the transcripts, the real meetings were almost unintelligible. The language contained a large number of unfinished sentences, false starts, overlapping utterances, interruptions and fillers, such as 'um', 'er' and 'you know'. A large proportion of the language contained jokes, quips, repetitions and asides. Some of the sentences were not grammatically correct'* (Williams, 1988, p. 45-58).

6. CONCLUSION

The computer allows researchers interested in the analysis of texts (literary or not) to add a valuable quantitative tool to their analyses. As Fortier (1995) writes, *'Using different software, specialists can obtain tables, graphs and statistics on the words that make up the works they are studying, as well as on syllables, punctuation marks, vocabulary, themes and lexical fields, etc. These results can then be used to compare authors or texts with each other. In particular, they can help the researcher to determine the authorship of a text, to distinguish imitations from authentic works, to understand the rhythmic patterns in verses or to grasp how a given author has contributed to the evolution of language'* (Fortier, 1995, p. 108).

Stylometry has mainly been used in literature and author identification by analysing writing style and use of lexis. As this paper suggests, it can also be used to identify styles of communication in electronic written professional discourse, differentiate between the vocabulary and expressions used by different organisations and professions, distinguish between formal and informal communication styles, thereby showing respect and building good relations and analysing and updating changes in the language used.

Automated text processing using algorithms should never be seen as anything other than a simple tool, which, while aiding the researcher, is by no means sufficient on its own. The researcher must indeed intervene before and after the automated analysis. Computer-assisted text analysis is not universally accepted, and experts have to contend with the limitations of this method.

Computer text analysis is marginalised by literary people. Most of them do not believe that computer tools can offer them any real help in their work and do not seem to have the curiosity to discover the possibilities of this tool. Unfortunately, a large number of articles dealing with automatic data processing in the field of computer text analysis are quite technical and sometimes off-putting for those who are not very familiar with statistics and computers. The human brain is able to grasp the meaning behind expressions, make associations of ideas, create and interpret metaphors, and even find new meanings in words that make up everyday vocabulary. No computer, even armed with the most advanced software, will ever be able to compete in intelligence and insight with a human author or reader.

Volume 5 Issue 2, 2021, pp. 51-61 doi: 10.22363/2521-442X-2021-5-2-51-61

References

Abbasi, A., & Chen, H. (2008). Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems, 26*(2), Article 7. https://doi.org/10.1145/1344411.1344413

Bailey, R. W. (1969). Statistics and style: A historical survey. In L. Doložel & R. W. Bailey (Eds.), *Statistics and style* (pp. 217-236). Elsevier.

Brunet, É. (2003). Peut-on mesurer la distance entre deux textes? *Corpus, 2,* 47-70. https://doi.org/10.4000/corpus.30

Burrows, J. F., & Craig, D. H. (1994). Lyrical drama and the 'turbid mountebanks': Styles of dialogue in romantic and renaissance tragedy. *Computers and the Humanities*, *28*(2), 63-86. https://doi.org/10.1007/BF01830688

Chaski, C. E. (2005). Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence, 4*(1), 1-13.

Chen, X., Hao, P., Chandramouli, R., & Subbalakshmi, K. P. (2011). Authorship similarity detection from email messages. In P. Perner (Ed.), *Machine learning and data mining in pattern recognition* (pp. 375-386). Springer. https://doi.org/10.1007/978-3-642-23199-5_28

Cheng, N., Chandramouli, R., & Subbalakshmi, K. P. (2011). Author gender identification from text. *Digital Investigation*, *8*(1), 78-88. https://doi.org/10.1016/j.diin.2011.04.002

Cho, K. H., Raiko, T., & Ilin, A. (2013). Gaussian-bernoulli deep boltzmann machine. In *Proceedings of the 2013 International Joint Conference on Neural Networks* (pp. 1-7). IEEE. doi: https://doi.org/10.1109/IJCNN.2013.6706831

Eder, M. (2017). Visualization in stylometry: Cluster analysis using networks. *Digital Scholarship in the Humanities, 32*(1), 50-64. http://dx.doi.org/10.1093/llc/fqv061

Fortier, P. A. (1995). Categories, theory, and words in literary texts. In G. Perissinotto (Ed.), *Research in Humanities Computing* (5th ed., pp. 91-109). Clarendon Press.

Hadjidj, R., Debbabi, M., Lounis, H., Iqbal, F., Szporer, A., & Benredjem, D. (2009). Towards an integrated e-mail forensic analysis framework. *Digital Investigation*, *5*(3-4), 124-137. https://doi.org/10.1016/j.diin.2009.01.004

Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities*, *28*(2), 87-106. https://doi.org/10.1007/BF01830689

Holmes, D. I. (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, *13*(3), 111-117. https://doi.org/10.1093/llc/13.3.111

Holmes, D. I. (2004). Stylometry. In S. Kotz, C. B. Read, N. Balakrishnan, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences* (2nd ed., pp. 378-386). Wiley-Blackwell Publishing House Ltd.

Kucukyilmas, T., Cambasoglu, B. B., Aykanat, C., & Can, F. (2008). Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing & Management*, *44*(4), 1448-1466. https://doi.org/10.1016/j.ipm.2007.12.009

Laffal, J. (1995). A concept analysis of Jonathan Swift's 'A tale of a Tub' and 'Gulliver's Travels'. *Computers and the Humanities*, *29*(5), 339-361. https://doi.org/10.1007/BF02279526

Lowe, D., & Matthews, R. (1995). Shakespeare vs. Fletcher: A stylometric analysis by radial basis functions. *Computers and the Humanities*, *29*(6), 449-461. https://doi.org/10.1007/BF01829876

Mendenhall, T. (1888). The characteristic curves of composition. *Science, 9*(214S), 237-246. https://doi.org/10.1126/science.ns-9.214S.237

Miall, D. S. (1995). Anticipation and feeling in literary response: A neuropsychological perspective. *Poetics*, *23*(4), 275-298. https://doi.org/10.1016/0304-422X(95)00004-4

Olsen, M. (1993). Signs, symbols and discourses: A new direction for computer-aided literature studies. *Computers and the Humanities*, *27*(5-6), 309-314. https://doi.org/10.1007/BF01829380

Ormerod, B., Volet, J. M., & Jaccomard, H. (1994). The female voice and traditional discourse biases: The case of francophone African literature. *Computers and the Humanities*, *28*(6), 353-367. https://doi.org/10.1007/BF01829970

Potter, R. G. (Ed.). (1989). *Literary computing and literary criticism: Theoretical and practical essays on theme and rhetoric*. University of Pennsylvania Publishing Press & Wiley-Blackwell Publishing House Ltd.

Savoy, J. (2020). *Machine learning methods for stylometry: Authorship attribution and author profiling*. Springer International Publishing. https://doi.org/10.1007/978-3-030-53360-1

Schuster, T., Schuster, R., Shah, D. J., & Barzilay, R. (2020). The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics, 46*(2), 499-510.

Sigelman, L., & Jacoby, W. (1996). The not-so-simple art of imitation: Pastiche, literary style, and Raymond Chandler. *Computers and the Humanities*, *30*(1), 11-28. https://doi.org/10.1007/BF00054025

Vartanova, E., Anikina, M., Dunas, D., & Gureeva, A. (2020). Media studies in Russia: Determination of scientific status. *Russian Journal of Communication, 12*(1), 1-15. https://doi.org/10.1080/19409419.2020.1729456

Williams, M. (1988). Language taught for meetings and language used in meetings: Is there anything in common? *Applied Linguistics, 9*(1), 45-58.

Zvereva, V. (2021). Social media users in search of 'facts': The Trade Union House fire case. *Russian Journal of Communication, 13*(1), 11-28. https://doi.org/10.1080/19409419.2021.1893212

JEAN LANGLOIS
Sciences Po | 23, Rue St. Guillaume, 75337 Paris, France Cedex 07
jean.langlois@sciencespo.fr