

# Function Words in Authorship Attribution From Black Magic to Theory?

Mike Kestemont

University of Antwerp

CLiPS Computational Linguistics Group

Prinsstraat 13, D.188

B-2000, Antwerp

Belgium

mike.kestemont@uantwerpen.be

## Abstract

This position paper focuses on the use of function words in computational authorship attribution. Although recently there have been multiple successful applications of authorship attribution, the field is not particularly good at the explication of methods and theoretical issues, which might eventually compromise the acceptance of new research results in the traditional humanities community. I wish to partially help remedy this lack of explication and theory, by contributing a theoretical discussion on the use of function words in stylometry. I will concisely survey the attractiveness of function words in stylometry and relate them to the use of character n-grams. At the end of this paper, I will propose to replace the term ‘function word’ by the term ‘functor’ in stylometry, due to multiple theoretical considerations.

## 1 Introduction

Computational authorship attribution is a popular application in current stylometry, the computational study of writing style. While there have been significant advances recently, it has been noticed that the field is not particularly good at the explication of methods, let alone at developing a generally accepted theoretical framework (Craig, 1999; Daelemans, 2013). Much of the research in the field is dominated by an ‘an engineering perspective’: if a certain attribution technique performs well, many researchers do not bother to explain or interpret this from a theoretical perspective. Thus, many methods and procedures continue to function as a black box, a situation which might eventually compromise the acceptance of experimental results (e.g. new attributions) by scholars in the traditional humanities community.

In this short essay I wish to try to help partially remedy this lack of theoretical explication, by contributing a focused theoretical discussion on the use of function words in stylometry. While these features are extremely popular in present-day research, few studies explicitly address the methodological implications of using this word category. I will concisely survey the use of function words in stylometry and render more explicit why this word category is so attractive when it comes to authorship attribution. I will deliberately use a generic language that is equally intelligible to people in linguistic as well as literary studies. Due to multiple considerations, I will argue at the end of this paper that it might be better to replace the term ‘function word’ by the term ‘functor’ in stylometry.

## 2 Seminal Work

Until recently, scholars agreed on the supremacy of word-level features in computational authorship studies. In a 1994 overview paper Holmes (1994, p. 87) claimed that ‘to date, no stylometrist has managed to establish a methodology which is better able to capture the style of a text than that based on lexical items’. Important in this respect is a line of research initiated by Mosteller and Wallace (1964), whose work marks the onset of so-called non-traditional authorship studies (Holmes, 1994; Holmes, 1998). Their work can be contrasted with the earlier philological practice of authorship attribution (Love, 2002), often characterized by a lack of a clearly defined methodological framework. Scholars adopted widely diverging attribution methodologies, the quality of whose results remained difficult to assess in the absence of a scientific consensus about a best practice (Stamatatos, 2009; Luyckx, 2010). Generally speaking, scholars’ subjective intuitions (*Gelehrtenintuition, connoisseurship*) played far too large a role and the low level of methodological explicitness in

early (e.g. nineteenth century) style-based authorship studies firmly contrasts with today's prevailing criteria for scientific research, such as replicability or transparency.

Apart from the rigorous quantification Mosteller and Wallace pursued, their work is often praised because of a specific methodological novelty they introduced: the emphasis on so-called function words. Earlier authorship attribution was often based on checklists of stylistic features, which scholars extracted from known oeuvres. Based on their previous reading experiences, expert readers tried to collect style markers that struck them as typical for an oeuvre. The attribution of works of unclear provenance would then happen through a comparison of this text's style to an author's checklist (Love, 2002, p. 185–193). The checklists were of course hand-tailored and often only covered a limited set of style markers, in which lexical features were for instance freely mixed with hardly comparable syntactic features. Because the checklist's construction was rarely documented, it seemed a matter of scholarly taste which features were included in the list, while it remained unclear why others were absent from it.

Moreover, exactly because these lists were hand-selected, they were dominated by striking stylistic features that because of their low overall frequency seemed whimsicalities to the human expert. Such low-frequency features (e.g. an uncommon noun) are problematic in authorship studies, since they are often tied to a specific genre or topic. If such a characteristic was absent in an anonymous text, it did not necessarily argue against a writer's authorship in whose other texts (perhaps in different topics or genres) the characteristic did prominently feature. Apart from the limited scalability of such style (Luyckx, 2010; Luyckx and Daelemans, 2011), a far more troublesome issue is associated with them. Because of their whimsical nature these low-frequency phenomena could have struck an author's imitators or followers as strongly as they could have struck a scholar. When trying to imitate someone's style (e.g. within the same stylistic school), those low-frequency features are the first to copy in the eyes of forgers (Love, 2002, p. 185–193). The fundamental novelty of the work by Mosteller and Wallace was that they advised to move away from a language's low-frequency features to a language's

high-frequency features, which often tend to be function words.

### 3 Content vs Function

Let us briefly review why function words are interesting in authorship attribution. In present-day linguistics, two main categories of words are commonly distinguished (Morrow, 1986, p. 423). The open-class category includes content words, such as nouns, adjectives or verbs (Clark and Clark, 1977). This class is typically large – there are many nouns – and easy to expand – new nouns are introduced every day. The closed-class category of function words refers to a set of words (prepositions, particles, determiners) that is much smaller and far more difficult to expand – it is hard to invent a new preposition. Words from the open class can be meaningful in isolation because of their straightforward semantics (e.g. 'cat'). Function words, however, are heavily grammaticalized and often do not carry a lot of meaning in isolation (e.g. 'the'). Although the set of distinct function words is far smaller than the set of open-class words, function words are far more frequently used than content words (Zipf, 1949). Consequently, less than 0.04% of our vocabulary accounts for over half of the words we actually use in daily speech (Chung et al., 2007, p. 347). Function words have methodological advantages in the study of authorial style (Binongo, 2003, p. 11), for instance:

- All authors writing in the same language and period are bound to use the very same function words. Function words are therefore a reliable base for textual comparison;
- Their high frequency makes them interesting from a quantitative point of view, since we have many observations for them;
- The use of function words is not strongly affected by a text's topic or genre: the use of the article 'the', for instance, is unlikely to be influenced by a text's topic.
- The use of function words seems less under an author's conscious control during the writing process.

Any (dis)similarities between texts regarding function words are therefore relatively content-independent and can be far more easily associated

with authorship than topic-specific stylistics. The underlying idea behind the use of function words for authorship attribution is seemingly contradictory: we look for (dis)similarities between texts that have been reduced to a number of features in which texts should not differ at all (Juola, 2006, p. 264–65).

Nevertheless, it is dangerous to blindly overestimate the degree of content-independence of function words. A number of studies have shown that function words, and especially (personal) pronouns, do correlate with genre, narrative perspective, an author's gender or even a text's topic (Herring and Paolillo, 2006; Biber et al., 2006; Newman et al., 2008). A classic reference in this respect is John Burrows's pioneering study of, amongst other topics, the use of function words in Jane Austen's novels (Burrows, 1987). This explains why many studies into authorship will in fact perform so-called 'pronoun culling' or the automated deletion of (personal) pronouns which seem too heavily connected to a text's narrative perspective or genre. Numerous empirical studies have nevertheless demonstrated that various analyses restricted to higher frequency strata, yield reliable indications about a text's authorship (Argamon and Levitan, 2005; Stamatatos, 2009; Koppel et al., 2009).

It has been noted that the switch from content words to function words in authorship attribution studies has an interesting historic parallel in art-historic research (Kestemont et al., 2012). Many paintings have survived anonymously as well, hence the large-scale research into the attribution of them. Giovanni Morelli (1816-1891) was among the first to suggest that the attribution of, for instance, a *Quattrocento* painting to some Italian master, could not happen based on 'content' (Wollheim, 1972, p. 177ff). What kind of coat Mary Magdalene was wearing or the particular depiction of Christ in a crucifixion scene seemed all too much dictated by a patron's taste, contemporary trends or stylistic influences. Morelli thought it better to restrict an authorship analysis to discrete details such as ears, hands and feet: such fairly functional elements are naturally very frequent in nearly all paintings, because they are to some extent content-independent. It is an interesting illustration of the surplus value of function words in stylometry that the study of authorial style in art history should depart from the ears,

hands and feet in a painting – its inconspicuous function words, so to speak.

#### 4 Subconsciousness

Recall the last advantage listed above: the argument is often raised that the use of these words would not be under an author's conscious control during the writing process (Stamatatos, 2009; Binongo, 2003; Argamon and Levitan, 2005; Peng et al., 2003). This would indeed help to explain why function words might act as an author invariant throughout an oeuvre (Koppel et al., 2009, p. 11). Moreover, from a methodological point of view, this would have to be true for forgers and imitators as well, hence, rendering function words resistant to stylistic imitation and forgery. Surprisingly, this claim is rarely backed up by scholarly references in the stylometric literature – an exception seems Koppel et al. (2009, p. 11) with a concise reference to Chung et al. (2007). Nevertheless, some attractive references in this respect can be found in psycholinguistic literature. Interesting is the experiment in which people have to quickly count how often the letter 'f' occurs in the following sentence:

Finished files are the result  
of years of scientific study  
combined with the experience  
of many years.

It is common for most people to spot only four or five instances of all six occurrences of the grapheme (Schindler, 1978). Readers commonly miss the *fs* in the preposition 'of' in the sentence. This is consistent with other reading research showing that readers have more difficulties in spotting spelling errors in function words than in content words (Drewnowski and Healy, 1977). A similar effect is associated with phrases like 'Paris in the the spring' (Aronoff and Fudeman, 2005, p. 40–41). Experiments have demonstrated that during their initial reading, many people will not be aware of the duplication of the article 'the'. Readers typically fail to spot such errors because they take the use of function words for granted – note that this effect would be absent for 'Paris in the spring spring', in which a content word is wrongly duplicated. Such a subconscious attitude needs not imply that function words would be unimportant in written communication. Con-

sider the following passage:<sup>1</sup>

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mtttaer in waht oreodr the ltteers in a wrod are, the olny iprmoentn tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

Although the words' letters in this passage seem randomly jumbled, the text is still relatively readable (Rawlinson, 1976). As the quote playfully states itself, it is vital in this respect that the first and final letter of each word are not moved – and, depending on the language, this is in fact not the only rule that must be obeyed. It is crucial however that this limitation causes the shorter function words in running English text to remain fairly intact (McCusker et al., 1981). The intact nature alone of the function words in such jumbled text, in fact greatly adds to the readability of such passages. Thus, while function words are vital to structure linguistic information in our communication (Morrow, 1986), psycholinguistic research suggests that they do not attract attention to themselves in the same way as content words do.

Unfortunately, it should be stressed that all references discussed in this section are limited to reader's experience, and not writer's experience. While there will exist similarities between a language user's perception and production of function words, it cannot be ruled out that writers will take on a much more conscious attitude towards function words than readers. Nevertheless, the apparent inattentiveness with which readers approach function words might be reminiscent of a writer's attitude towards them, although much more research would be needed in order to properly substantiate this hypothesis.

## 5 Character N-grams

Recall Holmes's 1994 claim that 'to date, no stylometrist has managed to establish a methodology which is better able to capture the style of

a text than that based on lexical items' (Holmes, 1994, p. 87). In 1994 other types of style markers (e.g. syntactical) were – in isolation – never able to outperform lexical style markers (Van Halteren et al., 2005). Interestingly, advanced feature selection methods did not always outperform frequency-based selection methods, that plainly singled out function words (Argamon and Levitan, 2005; Stamatatos, 2009). The supremacy of function words was challenged, however, later in the 1990s when character n-grams came to the fore (Kjell, 1994). This representation was originally borrowed from the field of Information Retrieval where the technique had been used in automatic language identification. Instead of cutting texts up into words, this particular text representation segmented a text into a series of consecutive, partially overlapping groups of  $n$  characters. A first order n-gram model only considers so-called unigrams ( $n = 1$ ); a second order n-gram model considers bigrams ( $n = 2$ ), and so forth. Note that word boundaries are typically explicitly represented: for instance, ' b', 'bi', 'ig', 'gr', 'ra', 'am', 'm '.

Since Kjell (1994), character n-grams have proven to be the best performing feature type in state-of-the-art authorship attribution (Juola, 2006), although at first sight, they might seem uninformative and meaningless. Follow-up research learned that this outstanding performance was not only largely language independent but also fairly independent of the attribution algorithms used (Peng et al., 2003; Stamatatos, 2009; Koppel et al., 2009). The study of character n-grams for authorship attribution has since then significantly grown in popularity, however, mostly in the more technical literature where the technique originated. In these studies, performance issues play an important role, with researchers focusing on actual attribution accuracy in large corpora (Luyckx, 2010). This focus might help explain why, so far, few convincing attempts have been made to interpret the discriminatory qualities of characters n-grams, which is why their use (like function words) in stylometry can be likened to a sort of black magic. One explanation so far has been that these units tend to capture 'a bit of everything', being sensitive to both the content and form of a text (Houvardas and Stamatatos, 2006; Koppel et al., 2009; Stamatatos, 2009). One could wonder, however, whether such an answer does much more than reproducing the initial question:

<sup>1</sup>Matt Davis maintains an interesting website on this topic: <http://www.mrc-cbu.cam.ac.uk/people/matt.davis/Cmabrigde/>. I thank Bram Vandekerckhove for pointing out this website. The 'Cmabridge'-passage as well the 'of'-example have anonymously circulated on the Internet for quite a while.

Then why does it work? Moreover, Koppel et al. expressed words of caution regarding the caveats of character n-grams, since many of them ‘will be closely associated to particular content words and roots’ (Koppel et al., 2009, p. 13).

The reasons for this outstanding performance could partially be of a prosaic, information-theoretical nature, relating to the unit of stylistic measurement. Recall that function words are quantitatively interesting, at least partially because they are simply frequent in text. The more observations we have available per text, the more trustworthily one can represent it. Character n-grams push this idea even further, simply because texts by definition have more data points for character n-grams than for entire words (Stamatatos, 2009; Daelemans, 2013). Thus the mere number of observations, relatively larger for character n-grams than for function words, might account for their superiority from a purely quantitative perspective.

Nevertheless, more might be said on the topic. Rybicki & Eder (2011) report on a detailed comparative study of a well-known attribution technique, Burrows’s Delta. John Burrows is considered one of the godfathers of modern stylometry – D.I. Holmes (1994) ranked him alongside the pioneers Mosteller and Wallace. He introduced his influential Delta-technique in his famous Busa lecture (Burrows, 2002). Many subsequent discussions agree that Delta essentially is a fairly intuitive algorithm which generally achieves decent performance (Argamon, 2008), comparing texts on the basis of the frequencies of common function words. In their introductory review of Delta’s applications, Rybicki and Eder tackled the assumption of Delta’s language independence: following the work of Juola (2006, p. 269), they question the assumption ‘that the use of methods relying on the most frequent words in a corpus should work just as well in other languages as it does in English’ (Rybicki and Eder, 2011, p. 315).

Their paper proves this assumption wrong, reporting on various, carefully set-up experiments with a corpus, comprising 7 languages (English, Polish, French, Latin, German, Hungarian and Italian). Although they consider other parameters (such as genre), their most interesting results concern language (Rybicki and Eder, 2011, p. 319–320):

while Delta is still the most successful method of authorship attribution based on word frequencies, its success is not independent of the lan-

guage of the texts studied. This has not been noticed so far for the simple reason that Delta studies have been done, in a great majority, on English-language prose. [...] The relatively poorer results for Latin and Polish, both highly inflected in comparison with English and German, suggests the degree of inflection as a possible factor. This would make sense in that the top strata of word frequency lists for languages with low inflection contain more uniform words, especially function words; as a result, the most frequent words in languages such as English are relatively more frequent than the most frequent words in agglutinative languages such as Latin.

Their point of criticism is obvious but vital: the restriction to function words for stylometric research seems sub-optimal for languages that make less use of function words. They suggest that this relatively recent discovery might be related to the fact that most of the seminal and influential work in authorship attribution has been carried out on English-language texts.

English is a typical example of a language that does not make extensive use of case endings or other forms of inflection (Sapir, 1921, chapter VI). Such weakly inflected languages express a lot of their functional linguistic information through the use of small function words, such as prepositions (e.g. ‘with a sword’). Structural information in these languages tends to be expressed through minimal units of meaning or grammatical morphemes, which are typically realized as individual words (Morrow, 1986). At this point, it makes sense to contrast English with another major historical lingua franca but one that has received far less stylometric attention: Latin.

Latin is a school book example of a heavily inflected language, like Polish, that makes far more extensive use of affixes: endings that which are added to words to mark their grammatical function in a sentence. An example: in the Latin word *ensi* (ablative singular: ‘with a sword’) the case ending (-i) is a separate morpheme that takes on grammatical role which is similar to that of the English preposition ‘with’. Nevertheless, it is not realized as a separate word separated by whitespace from surrounding morphemes. It is rather concatenated to another morpheme (ens-) expressing a more tangible meaning.

This situation renders a straightforward application of the Delta-method – so heavily biased towards words – problematic for more synthetic or agglutinative languages. What has been said about function words in previous stylometric research,

obviously relates to their special status as functional linguistic items. The inter-related characteristics of ‘high frequency’, ‘content-independence’ and ‘good dispersion’ (Kestemont et al., 2012) even only apply to them, insofar as they are grammatical morphemes. Luckily for English, a lot of grammatical morphemes can easily be detected by splitting running text into units that do not contain whitespace or punctuation and selecting the most frequent items among them (Burrows, 2002; Stamatatos, 2009). For languages that display another linguistic logic, however, the situation is far more complicated, because the functional information contained in grammatical morphemes is more difficult to gain access to, since these need not be solely or even primarily realized as separate words. If one restricts analyses to high-frequency words in these languages, one obviously ignores a lot of the functional information inside less frequent words (e.g. inflection).

## 6 Functors

At the risk of being accused of quibbling about terms, I wish to argue that the common emphasis on function words in stylometry should be replaced by an emphasis on the broader concept of functors, a term which can be borrowed from psycholinguistics, used to denote grammatical morphemes (Kwon, 2005, p. 1–2) or:

forms that do not, in any simple way, make reference. They mark grammatical structures and carry subtle modulatory meanings. The word classes or parts of speech involved (inflections, auxiliary verbs, articles, prepositions, and conjunctions) all have few members and do not readily admit new members (Brown, 1973, p. 75).

In my opinion, the introduction of the term ‘functor’ would have a number of advantages – the first and least important of which is that it is aesthetically more pleasing than the identical term ‘grammatical morphemes’. Note, first of all, that function words – grammatical morphemes realized as individual words – are included in the definition of a functor. The concept of a functor as such does not replace the interest in function words but rather broadens it and extends it towards all grammatical morphemes, whether they be realized as individual words or not. Note how all advantages, previously only associated with function words in stylometry (high frequency, good dispersion, content-independence, unconscious use) apply to every member in the category of functors.

A second advantage has to do with language independence. Note that stylometry’s ultimate goal regarding authorship seems of a universal nature: a majority of stylometrists in the end are concerned with the notorious Stylome-hypothesis (Van Halteren et al., 2005) or finding a way to characterize an author’s individual writing style, regardless of text variety, time and, especially, language. Restricting the extraction of functional information from text to the word level might work for English, but seems too language-specific a methodology to be operable in many other languages, as suggested by Rybicki and Eder (2011) and earlier Juola (2006, p. 269). Stylometric research into high-frequency, functional linguistic items should therefore break up words and harvest more and better information from text. The scope of stylistic focus should be broadened to include all functors.

The superior performance of character n-grams in capturing authorial style – in English, as well as other languages – seems relevant in this respect. First of all, the most frequent n-grams in a corpus often tend to be function words: ‘me’, ‘or’ and ‘to’ are very frequent function words in English, but they are also very frequent character bigrams. Researchers often restrict their text representation to the most frequent n-grams in a corpus (2009, p. 541), so that n-gram approaches include function words rather than exclude them. In addition, high-frequency n-grams are often able to capture more refined grammatical information. Note how a text representation in terms of n-grams subtly exploits the presence of whitespace. In most papers advocating the use of n-grams, whitespace is explicitly encoded. Again, this allows more observations-per-word but, in addition, makes a representation sensitive to e.g. inflectional information. A high frequency of the bigram ‘ed’ could reflect any use of the character series (reduce vs. talked). A trigram representation ‘ed ’ reveals a word-final position of the character series, thus indicating it being used for expressing grammatical information through affixation. Psycholinguistic research also stresses the important status of the first letter(s) of words, especially with respect to how words are cognitively accessed in the lexicon (Rubin, 1995, p. 74). Note that this word-initial aspect too is captured under an n-gram representation (‘ aspect’).

A widely accepted theoretical ground for the outstanding performance of character n-grams, will have to consider the fact that n-grams offer a more powerful way of capturing the functional information in text. They are sensitive to the internal morphemic structure of words, capturing many functors which are simply ignored in word-level approaches. Although some n-grams can indeed be ‘closely associated to particular content words and roots’ (Koppel et al., 2009, p. 13), I would be inclined to hypothesize that high-frequency n-grams work in spite of this, not because of this. This might suggest that extending techniques, like Delta, to all functors in text, instead of just function words, will increase both their performance and language independence.

A final advantage of the introduction of the concept of a functor is that it would facilitate the teaming up with a neighbouring field of research that seems extremely relevant for the field of stylometry from a theoretical perspective, but so far has only received limited attention in it: psycholinguistics. The many parallels with the reading research discussed above indicate that both fields might have a lot to learn from each other. An illustrative example is the study of functor acquisition by children. It has been suggested that similar functors are not only present in all languages of the world, but acquired by all children in an extremely similar ‘natural order’ (Kwon, 2005). This is intriguing given stylometry’s interest in the Stylome-hypothesis. If stylometry is ultimately looking for linguistic variables that are present in each individual’s *parole*, the universal aspects of functors further stress the benefits of the term’s introduction. All of this justifies the question whether the functor should not become a privileged area of study in future stylometric research.

## Acknowledgments

The author was funded as a postdoctoral research fellow by the Research Foundation of Flanders (FWO). The author would like to thank Matthew Munson, Bram Vandekerckhove, Dominiek Sandra, Stan Szpakowicz as well as the anonymous reviewers of this paper for their substantial feedback on earlier drafts. Finally, I am especially indebted to Walter Daelemans for the inspiring discussions on the topic of this paper.

## References

- S. Argamon and S. Levitan. 2005. Measuring the usefulness of function words for authorship attribution. In *Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing (2005)*. Association for Computing and the Humanities.
- S. Argamon. 2008. Interpreting Burrows’s Delta: Geometric and Probabilistic Foundations. *Literary and Linguistic Computing*, (23):131–147.
- M. Aronoff and K. Fudeman. 2005. *What is Morphology?* Blackwell.
- D. Biber, S. Conrad, and R. Reppen. 2006. *Corpus linguistics - Investigating language structure and use*. Cambridge University Press, 5 edition.
- J. Binongo. 2003. Who Wrote the 15th Book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, (16):9–17.
- R. Brown. 1973. *A First Language*. Harvard University Press.
- J. Burrows. 1987. *Computation into Criticism: A Study of Jane Austen’s Novels and an Experiment in Method*. Clarendon Press; Oxford University Press.
- J. Burrows. 2002. ‘Delta’: a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, (17):267–287.
- C. Chung and J. Pennebaker. 2007. The psychological functions of function words. In K. Fiedler et al., editor, *Social Communication*, pages 343–359. Psychology Press.
- H. Clark and E. Clark. 1977. *Psychology and language: an introduction to psycholinguistics*. Harcourt, Brace & Jovanovich.
- H. Craig. 1999. Authorial attribution and computational stylistics: if you can tell authors apart, have you learned anything about them? *Literary and Linguistic Computing*, 14(1):103–113.
- W. Daelemans. 2013. Explanation in Computational Stylometry. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 2, CICLing’13*, pages 451–462, Berlin, Heidelberg. Springer-Verlag.
- A. Drewnowski and A. Healy. 1977. Detection errors on the and and: Evidence for reading units larger than the word. *Memory & Cognition*, (5).
- S. Herring and John C. Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459.
- D. Holmes. 1994. Authorship Attribution. *Computers and the Humanities*, 28(2):87–106.

- D. Holmes. 1998. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, 13(3):111–117.
- J. Houvardas and E. Stamatatos. 2006. N-gram feature selection for authorship identification. In J. Euzenat and J. Domingue, editors, *Proceedings of Artificial Intelligence: Methodologies, Systems, and Applications (AIMSA 2006)*, pages 77–86. Springer-Verlag.
- P. Juola. 2006. Authorship Attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334.
- M. Kestemont, W. Daelemans, and D. Sandra. 2012. Robust Rhymes? The Stability of Authorial Style in Medieval Narratives. *Journal of Quantitative Linguistics*, 19(1):1–23.
- B. Kjell. 1994. Discrimination of authorship using visualization. *Information Processing and Management*, 30(1):141–50.
- M. Koppel, J. Schler, and S. Argamon. 2009. Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.
- E. Kwon. 2005. The Natural Order of Morpheme Acquisition: A Historical Survey and Discussion of Three Putative Determinants. *Teachers' College Columbia Working Papers in TESOL and Applied Linguistics*, 5(1):1–21.
- H. Love. 2002. *Authorship Attribution: An Introduction*. Cambridge University Press.
- K. Luyckx and W. Daelemans. 2011. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, (26):35–55.
- K. Luyckx. 2010. *Scalability Issues in Authorship Attribution*. Ph.D. thesis, University of Antwerp.
- L. McCusker, P. Gough, and R. Bias. 1981. Word recognition inside out and outside in. *Journal of Experimental Psychology: Human Perception and Performance*, 7(3):538–551.
- D. Morrow. 1986. Grammatical morphemes and conceptual structure in discourse processing. *Cognitive Science*, 10(4):423–455.
- F. Mosteller and D. Wallace. 1964. *Inference and disputed authorship: The Federalist*. Addison-Wesley.
- M. Newman, C. Groom, L. Handelman, and J. Pennebaker. 2008. Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes*, 45(3):211–236, May.
- F. Peng, D. Schuurmans, V. Keselj, and S. Wang. 2003. Language independent authorship attribution using character level language models. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 267–274.
- D. Rubin. 1995. *Memory in Oral Traditions. The Cognitive Psychology of Epic, Ballads and Counting-out Rhymes*. Oxford University Press.
- J. Rybicki and M. Eder. 2011. Deeper Delta across genres and languages: do we really need the most frequent words? *Literary and Linguistic Computing*, pages 315–321.
- E. Sapir. 1921. *Language: An Introduction to the Study of Speech*. Harcourt, Brace & Co.
- R. Schindler. 1978. The effect of prose context on visual search for letters. *Memory & Cognition*, (6):124–130.
- E. Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society For Information Science and Technology*, (60):538–556.
- H. Van Halteren, H. Baayen, F. Tweedie, M. Haverkort, and A. Neijt. 2005. New Machine Learning Methods Demonstrate the Existence of a Human Stylome. *Journal of Quantitative Linguistics*, (12):65–77.
- R. Wollheim. 1972. *On Art and the Mind: Essays and Lectures*. Harvard University Press.
- G. Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley.