The Exact Analysis of Text

John Nerbonne,* University of Groningen

July, 2007

Introduction

More than forty years after its initial publication, Frederick Mosteller and David Wallace's *Inference and Disputed Authorship: The Federalist* is an excellent introduction to the application of exact techniques to various problems in the analysis of text; in some respects it is still unequaled. Given the way the computer has revolutionized exact textual analysis in the last thirty years, and in particular in light of the explosion in the availability of digital texts in the last ten, it is remarkable that any work could continue to contribute for so long. When the original Addison-Wesley edition went out of print, the book was republished in 1984 by Springer (with some emendations) under the title *Applied Bayesian and Classical Inference, the Case of the Federalist Papers*, which, however, has also gone out of print.. We all owe thanks to CSLI publications of Stanford and publisher Dikran Karaguezian for reissuing this classic monograph in their David Hume series, specialized in mathematical studies in the Humanities, making it available once again to scholars and students. At Prof. Wallace's request, this republication bears the original, more fitting title once again.

Google Scholar lists several hundred citations for *Inference and Disputed Authorship*, most of which date since 1980; this is surely enough to demonstrate the continued impact of the book. But there are many explicit recommendations as well. In a survey article on Bayesian techniques in information retrieval in 1998 David Lewis recommended that "the book by Mosteller and Wallace is the most clear treatment from a classification point of view" (Lewis, 1998, p.7) (referring to the treatment of underlying word distributions). Mosteller and Wallace's (1964) work is also used extensively by computational linguists (Church and Mercer, 1993, Hindle and Rooth, 1993, Yarowsky, 1995), but most of all in authorship studies.

Holmes and Forsyth (1995) surveys statistical authorship studies using the publication of *Inference and Disputed Authorship* as the watershed from which

*I am pleased to acknowledge the support of the *Musée de l'Homme, Hommes Natures Société*, CNRS UMR 5145, where I was a guest while writing this introduction. Thanks too to Jerom Janssen for reactions to an early version.

they begin. Joseph Rudman says of authorship studies in general (Rudman, 2002):

> The study that arguably is the most famous and the most successful is the Mosteller and Wallace work on the twelve disputed Federalist papers. [...] Almost every non-traditional authorship study [...] cites Mosteller and Wallace for one reason or another.

In the remainder of this brief introduction we review some of the ways in which Mosteller & Wallace's book has been influential, especially in popularizing Bayesian analysis, in indicating how texts may be classified according to various criteria, and, in particular, how authorship may be inferred.

Bayesian Analysis

We first sketch very roughly the basic ideas of Bayesian analysis. Our hope is to capture the interest of those who might other avoid this topic (or this book) as too technical. This is the background provided to introductory students in *Humanities Computing* at the University of Groningen to interest them in text analysis using Bayesian techniques. The basic ideas are genuinely simple, as we shall see.

Thomas Bayes (1702-1761) studied conditional probability of the sort we are confronted with daily, for example, when we read that young drivers are more likely to violate traffic laws than older ones, that the religious tend to be politically conservative, or that better educated people live longer (we ignore whether the claims are true). To see the simplicity of Bayes's theorem, we use the usual symbolism, $P(A|B)$ to refer to the probability of A given B. We might then write $P(L|E)$ to refer to the chance of living long L (say, past 85) given that one's education level is high, E (say college-level or more). The natural view of conditional probability, $P(A|B)$, defines it as the proportion of probability reserved for the co-occurrence (joint occurrence) of A and B as compared to the probability of B:

$$P(A|B) = P(A,B)/P(B),$$

where $P(A,B)$ denotes the probability of A and B occurring jointly. Continuing our example, when we speak of the chance of living long given that one's education level is high, $P(L|E)$, we proceed from the probability of being highly educated, and then compare the fraction of those highly educated who live long. Saying this another way, we compare the chance of having a college degree and living past the age of 85 to the chance of having a college degree in general (no matter what life span). To see whether the chance of living long is *better* if one is well educated, we should ask whether $P(L|E) > P(L|\overline{E})$, where the latter, $P(L|\overline{E})$, is the chance of living long if one is *not* well educated ($\overline{E}$). We hope that the symbols begin to feel familiar.

Bayes noticed that there is an intimate relation between $P(A|B)$ and $P(B|A)$, a circumstance which has come to be known as BAYESIAN INVERSION. By simple algebra we see that, given $P(A|B) = P(A,B)/P(B)$ (def.), it must also be true that $P(A|B) \cdot P(B) = P(A,B)$. By the same manipulation we proceed from $P(B|A) = P(B,A)/P(A)$ (def.) to $P(B|A) \cdot P(A) = P(B,A)$. Since $P(A,B)$ and $P(B,A)$ both denote the chance of the joint occurrence of A and B, they are necessarily the same, so that we may conclude that $P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$. Dividing both sides by $P(B)$ gives us Bayes's theorem (or Bayes's law):

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

The inversion is critical in cases where it is easier to obtain information about the one conditional probability than the other. We continue immediately to this case.

Bayes's law is applied in Bayesian analysis to the situation where we reason about data D and hypotheses h. In this case we proceed from the following instantiation of the general law:

$$P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}$$

This formulation suggests the interest that Bayesian inversion has held for problems in scientific and technical inference. Reasoning from data to hypotheses is the soul of scientific inference, and Bayes's insight lets us operationalize some aspects of this in a useful fashion. In fact Bayesian analysis normally simplifies the basic step even further.

In the usual situation we are comparing several alternative hypotheses $h_1, \ldots, h_n$ and asking which is best. In the formulation above we ask which of $\{P(h_1|D) \ldots P(h_n|D)\}$ is greatest, a specification which is normally formulated $\text{argmax}_{h_i} P(h_i|D)$, i.e. for which argument h_i of the expression does $P(h_i|D)$ have its maximal value. Because we are normally examining only the single data set, the effect of $P(D)$ on the right-hand side can never decide which h_i leads to a maximal value, so that we can ignore it in searching for the optimal h. This leads us to the statement of Bayes's law as it is normally applied:

$$\text{argmax}_{h_i} P(h_i|D) = \text{argmax}_{h_i} P(D|h_i) \cdot P(h_i)$$

The left-hand side of the equation is simply the statement of the problem, for which the right-hand side provides (a recipe for) a solution. Note that there are two terms on the right, $P(D|h_i)$, the LIKELIHOOD of the data given the hypothesis, and $P(h_i)$, the PRIOR PROBABILITY of the hypothesis, also known simply as the PRIOR. Both are important.

One big reason for the appeal of Bayesian analysis is its undeniable respect for empirical facts. Calculating the likelihoods $P(D|h_i)$ is always an essential part of a Bayesian analysis, and this normally means collecting a great deal of data noting which h_i was at play, the present book being no exception. Let's examine this more concretely in order to clarify the way in which working with Bayesian likelihoods is "data driven". Mosteller & Wallace study the authorship of the *Federalist Papers*, a series of political essays written primarily by Alexander Hamilton and James Madison during the period in which the U.S. constitution was being considered for ratification. Each of the essays was written by one of the two but each was simply signed 'Publius', and there has been discussion as to which essays should rightfully be attributed to which man. For each essay, we are therefore asking which h_i, which we might in this case note as h_H and h_M (for Hamilton and Madison, respectively), is responsible for the data. For the purpose of the study, the aspect of the data that was kept in focus was the words used (we return to this below). We are searching for $\text{argmax}_{h_i} P(h_i|D)$, which will depend on a prior probability (which we ignore for the moment) but also on the likelihood $P(D|h_i)$. Recalling that we are estimating the relevant aspects of the data by keeping track of which words were used how often, we obtain distinct likelihood estimates for h_H and h_M by tracking differences in the counts of the relevant words in those essays for which authorship *is* known. So the Bayesian inversion has as a necessary step keeping track of when evidence (D) has been seen in cases in which the h_i is known. In the case at hand, counts needed to be made of the frequency of relevant words in the *Federalist* papers (and comparable material) for which authorship is known. These counts allow the estimates of the likelihoods $P(h_M|D)$ and $P(h_H|D)$. The need to attend to estimate empirical likelihoods gives Bayesian analysis its strong empiricist flavor.

The strong empiricist flavor can come to dominate in some studies, where probabilities are modeled very simply and are estimated by simple frequency counts. But the relation between probabilities and frequencies is more complex, since in general many statistical models may be compatible with a given set of frequencies, a fact Mosteller & Wallace take great pains to keep in their reasoning, examining alternative probability distributions (pp. 32ff, p. 66), especially mixtures of distributions, and including re-estimations where the assumptions seem questionable (pp. 87–88). If only all contemporary studies took such care at this point![1]

Although Mosteller & Wallace downplay the role of the prior (their "initial odds") in determining the authorship of the *The Federalist Papers* (p. 56, p. 264), they acknowledge its importance, which can come to dominate in other areas in which Bayesian analysis has become popular. For example, Bayesian analysis is popular in medicine, where it is seen as part of the movement for "evidence-based medicine" (Guyatt et al., 1992). In medical applications hypotheses are diagnoses which are compared in their ability to explain the data, i.e. the symptoms shown by a patient. When it comes to inspecting the inversions, we may find that several diagnoses could predict the existing symptoms, in which cases the priors can easily dominate, suggesting that the diagnosis should focus on the most frequently occurring illnesses.

It is worth adding one further point to this introductory presentation about

[1] Modeling word frequency distributions continues to be problematic. Baayen (2001) recently proposed new models of word frequency distributions which he dubs "Large Number of Rare Event" distributions, which seem promising, but there is still no scholarly consensus about the exact form of linguistic frequency distributions.

the role of assumptions of independence, which we have ignored until now. Mosteller & Wallace calculate the probabilities of words occurring in documents as if the words occurred independently of one another (pp. 35–37, p. 111, p. 115). This allows them to combine chances with respect to hypotheses by simply multiplying the chances of seeing the words used. The problem is that words do not co-occur in a statistically independent way, at least not in general. Mosteller & Wallace are aware of this, and, again, in honorable exception to many subsequent uses of Bayesian analysis, they examine the degree to which this assumption of independence is violated, and attempt to reestimate, correcting for the (few) points at which it fails in their data (p. 84). Mosteller & Wallace do not use the term NAIVE BAYES to describe an application relying on this independence assumption, but Bayesian analysis in which one assumes that the aspects of the data ones is using for classification are statistically independent has come to be known as NAIVE BAYES. If the term was invented with intent to add pejorative flavor, it has long since lost that.

All in all, a paradigm example of the applications of Bayesian techniques, and, one which succeeded in contributing to the historical debate about the authorship of an important collection of essays.

Bayesian Fortunes

Mosteller & Wallace's success in applying Bayesian analysis did not go unnoticed, and was part of a major increase in the popularity of this sort of work. There are now even textbooks introducing statistical analysis via Bayes rather than via the classical, hypothesis testing mode (Berry, 1996, Albert and Rossman, 2001).

The authors emphasize as well the philosophical advantage for Bayesians in applying statistics to judgments about single events. The alternative is to regard statistical propositions as FREQUENTIST, in which case one seeks to identify classes of events whose frequency distributions may be studied. Concrete propositions attributing authorship are so specific, however, that this interpretation is counterintuitive. Hamilton either did or did not write Federalist paper Nr. 52, to use Mosteller & Wallace's example, and there's little insight gained by considering a class of such events. There was either one such event or none at all.

Bayesian models are used very frequently in (computational) linguistics. One popular application is the disambiguation of words such as 'party', which can be a festive gathering, a political organization, someone involved in a legal procedure or arrangement such as a contract, or a predicate meaning something like 'a participant, as in 'he wouldn't be party to the plan'. The hypotheses are these different meanings, and the data is taken to be neighboring words and their syntactic parts of speech, for which likelihoods can be tallied (Yarowsky, 1995). But given how frequently linguists wish to detect latent structure in texts, it is not surprising that Bayesian techniques are part of their standard tool box. And there are further applications that go beyond the detection of latent linguistic structure; Ellison (2007) uses a Baysian model to draw inferences about earlier language stages, e.g. the existence of sound changes and the fact that words in different languages may be similar in form and meanings due to the fact that they are COGNATE, i.e. descendants of a single word in an earlier historical stage of the language. In Ellison's case the hypotheses are the forms of the word(s) at earlier stages and also the sound changes that lead to new pronunciations, the data are the forms at later stages, and his model compares the chances of arriving at the later forms. Fortuitous overlap (misleadingly similar words) can be compared in likelihood to overlap due to cognacy. There are many other examples in which Baysian reasoning is invoked in contemporary language analysis.

Text Classification

Inference and Disputed Authorship is devoted to a study in authorship attribution, which Mosteller & Wallace in turn construe as a text classification problem. Surely one of the reasons for the continued high level of interest in the book is the enormous growth in interest in text classification. As Mosteller & Wallace are careful to point out, the classes for the purposes of one classification problem need not be same as the classes for another.

INFORMATION RETRIEVAL (IR) studies the problem of finding documents that are relevant to the queries of information seekers (Salton and McGill, 1986).[2] The document classes are therefore no longer defined by their authorship, but rather by their relevance to users' expressed informational needs. While IR was a fairly small field before the advent of the world-wide web, it is now a sizeable industry fed by the apparent impossibility of *imposing* structure on the web and users' insatiable wishes nonetheless to find their way in it. Bayesian techniques (including especially Naive Bayes) are part of the standard repertoire in (IR) (Lewis, 1998). In addition to using Bayesian techniques to estimate the relevance of the context of a text on the web, most effective systems (notably, Google) likewise include some way of estimating the "authority" of a web site based on the number (and "authority") of web sites which include a pointer to it. We do not wish to suggest that the problem of providing good access to public information is solved, only that Bayesian techniques are proving useful there.

In some ways a converse to the problem of seeking information, there is likewise a problem in excluding importunate information, in particular unwanted email, better known as SPAM. The problem is therefore the construction of a *spam filter*, and while the class of text has changed, the problem may still be construed as one of text classification, where the class definitions are now "email which the recipient would like to receive" and "email which is the unwanted product of mass mailing" (the intention is that these classes partition the set of incoming email so that each email belongs to exactly one class). Again

[2] A much improved instruction to information retrieval is scheduled to appear in 2008, *Introduction to Information Retrieval* by Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze, Cambridge University Press. As of this writing it is available at Schütze's web page. See `http://www-csli.stanford.edu/~schuetze/`

Bayesian and, in particular, Naive Bayes techniques have been used with considerable success in constructing spam filters (Sahami et al., 1998, Goodman and Heckerman, 2004), even if experts are not sanguine about the ultimate chances of warding off spam while a small percentage of mass mailing is in fact appreciated (judging by responses, including purchases), and spam broadcasters are able to analyze and circumvent the tactics used by filters.

There are many other applications of text classification, including tracking the mention of a company or public figure through "clippings" from news items, the creation of specialized portals of information such as custom-made newspapers, or portals for elementary school children, etc. Very similar techniques are now also used to classify music, initially for the purpose of retrieving titles and identifying information for users who seek the information based on knowing several measures of melody or a refrain (Brochu and de Freitas, 2003, Typke et al., 2005).

One of the most difficult text classification problems is identifying authorship, the focus of *Inference and Disputed Authorship*, to which we now turn.

Authorship

As we have noted, authorship attribution may be viewed as a special form of text classification in which the classes are essentially just the works of given authors. A key question in authorship attribution has been to determine what sorts of *evidence* might bear on determining authorship. Most work on authorship has accepted two principles which Mosteller & Wallace introduced; first, one should search for evidence of authorship among high-frequency elements (Chap.2) even when there appear to be "markers", i.e. words that only one of the candidate authors appears to use (or where one candidate author uses the marker much more frequently). If the candidate marker is a high frequency word, such as *upon*, which Hamilton used with overwhelmingly greater frequency that Madison, then it can contribute soundly to the inference. But low frequency words are not encountered reliably enough to be of much use in authorship attribution (but see below). Holmes (1994:88) cites Bailey approvingly to the effect that the features studied for the purpose of authorship attribution should be "frequent and easily quantifiable and relatively immune from conscious control". Holmes goes to conclude (p.104) that researchers should be prepared to investigate different sets of variables. As a second point of contribution, we note that Mosteller & Wallace deliberately aggregated over many words of individually weak discriminative value, and subsequent research has followed them in this as well.

Not every researcher has used words as evidence of authorship. Kenny (1982) uses the part of speech of sentence-final words in classical Greek (see also Baayen et al. (1996) and Holmes (1994) for further suggestions about useful evidence other than lexical choice). But students of authorship attribution have largely followed Mosteller & Wallace's lead. John Burrows, one of the acknowledged leaders in the sub-discipline of exact text analysis, worked exclusively with high frequency words in his so-called "delta", an average standard (z-score) difference for a set of high-frequency words, accepting both parts of the Mosteller & Wallace heritage (Burrows, 2002, 2003). We return to Burrows below, summarizing that there has been remarkably widespread acceptance that high-frequency elements provide the most reliable clues, reflecting, as they do, unconscious tendencies of authors. It is striking that Mosteller & Wallace's conclusion still stands, but even more striking that exactly the same reasoning is offered after so many years (noting again the advice above of Bailey, quoted by Holmes).

The attribution of authorship in the case of literary texts regularly inspires new technical and scholarly work (Hoover, 2004, Rudman, 2002) as scholars attempt to tease apart e.g. effects of authorship from editorial effects, the effects of depicting characters, and authorial purpose (e.g., whether an author was reacting to a specific piece).

In recent work, motivated by the wish to find discriminating indications that would jibe better with literary intuition (than the analysis of distributions of highly frequent words), Burrows postulates that "Evidence of authorship pervades whatever anybody writes" (Burrows, 2007, p.28), proposing statistical techniques that function on less frequent words. This very recent paper is, of course, an exception to the general acceptance of the need to focus on high frequency words.

Authorship detection has not attracted new practitioners in the same large numbers as IR, but it is worth noting that applications other than authorship attribution for anonymous literary tests have also been developed, and these add research energy to the field. We close this section by noting several of these.

- The exposure of literary forgery is surely closely related to revealing the authors of anonymous works (Love, 2002, Ch. 10), but it is naturally complicated by the fact that forgers try deliberately to mislead. Love (2002:185), reporting on unpublished work by Burrows, is optimistic about detecting the unconscious speech habits of authors, i.e. of the sort found in the distributions of high-frequency words.

- Forensic linguistics studies language use for its value as legal evidence and as an aid in solving and preventing crime. While most use of authorship techniques are too uncertain for broad use in criminal or civil proceedings, there is interest in pursuing them for their value in solving crime (Chaski, 2005). Issues include the legitimacy of wills and the authorship of documents containing threats.

- Plagiarism may always have been a problem in schools and colleges, but it has become easier to commit and easier to detect since the advent of the internet. Normally plagiarizers copy so much text from unacknowledged sources that the question of statistical inference pales before the question of locating the putative source. Once the source is found, the proof is often overwhelming Uzuner et al. (2005).[3]

[3]It is much more difficult to detect plagiarism if there is no putative source available for comparison, but this would also seem worthwhile, as it would provide an indication of when the search for an source would be worthwhile, but I do not know of work in this direction.

- Karina van Dalen-Oskam found a clever use for authorship attribution techniques in analyzing the medieval text *Walewein*, which is known to have been the work of two authors whose division of labor is uncertain. She was able to generate a new hypothesis about the point of division between the sections which were the work of the two writers (van Dalen-Oskam 2007).

Let us end this section by indicating where interested readers might best pursue interests in authorship attribution whetted by the republication of *Inference and Disputed Authorship*. Holmes (1994) surveys the statistics behind both Mosteller and Wallace's work but also statistical work on authorship attribution since then (what literary scholars, including Rudman (2002), sometimes refer to as "non-traditional authorship studies"). Love (2002) recommends itself for attention to the many sorts of evidence that have been used in authorship attribution and for an engaging survey of work since the 1960's. His book shows both the energy that has been invested in authorship studies since *Inference and Disputed Authorship* and especially the need to involve domain experts, i.e., literary scholars, in such studies. While neither Mosteller nor Wallace was an expert on eighteenth century language or political thought, we suspect from the great care that was taken in their book to incorporate insights from non-statistical studies that they might wholeheartedly agree with this development.

Conclusions and Prospects

The focused contribution of *Inference and Disputed Authorship* was to develop exact techniques for the analysis of texts, so that it is not surprising that contemporary textbooks devote entire chapters to Mosteller and Wallace's legacy to this field (Lebart and Salem, 1994). As we have noted above, the work has also stimulated further research in computational linguistics, particularly in the problem of disambiguating words; the techniques it championed have become part of the standard set of techniques used in text classification, which are now applied in information retrieval in SPAM detection and even in music retrieval; and even today, more than forty years after its original publication, it is perennially and positively cited in virtually all works on authorship attribution.

References

Albert, James and Allan Rossman. 2001. *Workshop Statistics: Discovery with Data, A Bayesian Approach.* New York and Berlin: Springer.

Baayen, R. Harald. 2001. *Word Frequency Distributions.* Dordrecht: Kluwer.

Baayen, R. Harald, Hans van Halteren, and Fiona Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11(3):121.

Berry, Donald. 1996. *Statistics: A Bayesian Perspective.* London: Duxbury.

Brochu, Eric and Nando de Freitas. 2003. "Name that song!": A probabilistic approach to querying on music and text. In *Advances in Neural Information Processing Systems*, vol. 15.

Burrows, John. 2002. 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* 17:267–287.

Burrows, John. 2003. Questions of authorship: Attribution and beyond. *Computers and the Humanities* 37:1–26.

Burrows, John. 2007. All the way through: Authorship in different frequency strata. *Literary and Linguistic Computing* 17:27–47.

Chaski, Carole. 2005. Who's at the keyboard? Authorship attribution in digital evidence investigations. *International Journal of Digital Evidence* 4(1). `http://www.utica.edu/academic/institutes/ecii/ijde/`.

Church, Kenneth and Robert Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics* 19(1):1–24.

Dalen-Oskam, Karina van. 2007. Kwantificeren van stijl. *Nederlandse Taal- & Letterkunde* 123:37–54.

Ellison, T. Mark. 2007. Bayesian identification of cognates and correspondences. In *Computing and Historical Phonology. 9th SIGMORPHON*. Shroudsburg, PA: ACL.

Goodman, Joshua and David Heckerman. 2004. Fighting spam with statistics. *Significance* 1(2):69–72.

Guyatt, Gordon, John Cairns, and D. Churchill. 1992. Evidence-based medicine. A new approach to the teaching of medicine. *Journal of the American Medical Association* 268:2420–2425.

Hindle, Don and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics* 19(1):103–120.

Holmes, David. 1994. Authorship attribution. *Computers and the Humanties* 28:87–106.

Holmes, David and Richard Forsyth. 1995. The Federalist revisited: New directions in authorship attribution. *Literary and Linguistic Computing* 10(2):111–127.

Hoover, David. 2004. Testing Burrow's Delta. *Literary and Linguistic Computing* 19:477–495.

Kenny, Anthony. 1982. *The Computation of Style: An Introduction to Statistics for Students of Literature and Humanities*. Oxford: Pergamon.

Lebart, Ludovic and André Salem. 1994. *Statistique Textuelle*. Paris: Dunod.

Lewis, David. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, vol. 1398, pages 4–15. Heidelberg: Springer.

Love, Harold. 2002. *Attributing Authorship: An Introduction*. Cambridge: Cambridge University Press.

Rudman, Joseph. 2002. Non-traditional authorship attribution studies in eighteenth century literature. *Jahrbuch für Computerphilologie* 4. `http://computerphilologie.uni-muenchen.de/jg02/rudman.html`.

Sahami, Mehran, Susan Dumais, David Heckerman, and Erik Horvitz. 1998. A Bayesian approach to filtering junk e-mail. In *Learning for Text Categorization: Papers from the 1998 Workshop*, vol. 62. AAAI Technical Report WS-98-05.

Salton, Gerald and Michael J. McGill. 1986. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Typke, Rainer, Frans Wiering, and Remco Veltkamp. 2005. A survey of music information retrieval systems. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, pages 153–160.

Uzuner, Özlem, Boris Katz, and Thade Nahnsen. 2005. Using syntactic information to identify plagiarism. In *Proceedings of the ACL Workshop on Educational Applications*, pages 37–44.

Yarowsky, David. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Conference of the Association for Computational Linguistics*, pages 189–196.