MEASURING STYLE: DOCUMENT ANALYSIS AND FORENSIC STYLOMETRY

Patrick Juola, Ph.D., CFC, Diplomate

Juola & Associates

256 W. Schwab Avenue

Munhall, PA 15210

Email : pjuola@juolaassoc.com

Abstract: Traditional document analysis can fail when there is no traditional document, as in blog posts, email, or Word files. We describe an emerging forensic discipline of “stylometry,” the analysis of writing style with an eye to identifying or profiling the writer of a document. We describe the theory, methods, strengths, and weaknesses of this important subfield of forensic science, with an eye to practical applications.

Keywords : Stylometry, text analysis, authorship, profiling

Learning Objectives:

1. What stylometry is and what type of problems it can address
2. How stylometric analysis is performed
3. What kind of data is needed for a stylometric analysis to be useful

THE ISSUE

When is a document not a document? Consider the following situations:

* A dead body is found inside a dorm room, with his computer on and a suicide note typed in Microsoft word. The investigating officer wants to know whether the note is genuine.
* An anonymous post on a blog reveals confidential business information about a company. The company wants to know who wrote the post, but the foreign ISP will not cooperate.
* An inappropriate email was sent from a person’s desktop computer, under her name. The person in question is known to have been at a meeting during that time. Who wrote the email?
* A corrupt foreign official is believed to have allowed an interested party to write a decision for him illicitly, a decision worth millions to that party. Is the decision in fact the work of the official, or of his alleged partner?

All of these situations share in common a question---who wrote this piece of text? --- and a complication that there is no actual physical document for a questioned document specialist to examine. The science of document examination is well-established and such specialists are capable of working apparent miracles based on the slope of a handwritten t or a worn A key on a typewriter. But when working with electronic documents, there are no worn keys, and one 12-point Helvetica font A looks exactly like any other. Not just similar, but identical. In this case, how can we decide whether or not the suicide note is genuine?

The traditional answer is to look, not at the document itself, but at the language contained in the document. Close reading, hopefully by an expert in language, may be able to provide cues to the authorship. Unfortunately, this kind of close reading is subjective, time-consuming, and difficult to validate, and often as likely to produce a situation of “dueling experts” taking opposite sides based on quite different notions of what these cues might be. Especially in the Daubert age, where expert testimony must be rooted in scientific principles, falsifiable, and with measurable error rate, it is difficult for experts to have realistic answers to questions like “how likely are you to be wrong?”

If, on the other hand, some notion of writing style could be objectively measured, it would be possible to quantify exactly the idea that “this document is likely to have come from this hand.” The modern science of stylometry seeks to do just that. By quantifying individualized aspects of language, aspects such as choice of synonyms, favored grammatical structures, and so forth, forensic text analysts can make objective judgments to answer questions such as the above. Better still, since much of the work can be performed by computer, it can be done quickly, efficiently, and at low cost while maintaining an even lower error rate due to lack of human fatigue and error.

THEORY OF STYLOMETRY

The theory behind stylometry is quite simple. Just as every person has their own fingerprints or genome, everyone has their own individual writing and speaking style. Obvious examples of this kind of style, familiar and useful even before the age of computers, are spelling errors. Wellman (1936), for instance, describes a case he won by noting that the word *touch* was misspelled in the same way in several different documents, and was able to validate those documents by showing that the particular mistake he identified was characteristic of one specific witness. (In particular, he was able to elicit that mistake on the witness stand by asking her to write a sentence containing *touch*.) But in the event that we have no such helpful mistakes, there are other individual quirks that can easily be used.

Vocabulary is one such quirk. Consider for yourself -- what is the name of the huge piece of furniture in front of your TV, the one that seats at least three people? Depending upon where you are from (and how old you are), that might be a *sofa*, a *couch*, a *settee*, a *divan,* a *chesterfield*, or a *davenport*. The rubber-soled shoes you wear to work out might be *sneakers*, *tennis shoes*, or *trainers*; Wikipedia (n.d.) lists at least eighteen regional synonyms for these things. Something even bigger than *large* could be *huge*, *gigantic*, *colossal*, *enormous*, and so forth. And, of course, we are all cognizant of idiosyncratic individuals with a pronounced tendency toward sesquipedalian lexical items – or, alternatively, we all know some people who use a lot of big words.

FIGURE 1 ABOUT HERE

A more interesting example comes from figure 1. Look at the picture and tell yourself where the fork is. The important word in your answer is not what you think it is -- it’s the preposition you used. Is the fork *to* the left, *on* the left, *at* the left, or something else?

Prepositions, like pronouns, conjunctions, and determiners, are examples of “function words.” Unlike content words such as nouns, verbs, and adjectives, these words tend to be common, short, and not very meaningful. (I dare you to try to define the word “of”.) Instead, they describe relationships between other words – structural information, not content. In part because of this, people tend not to pay much attention to them. In one famous experiment (Bransford et al., 1983), subjects were largely unable to notice or remember whether a sentence they had just memorized was “Three turtles rested on a floating log, and a fish swam beneath them” or “Three turtles rested on a floating log and a fish swam beneath it.” The difference between the pronoun *them* and the pronoun *it* just isn’t that salient, especially as, in the situation described, the fish would have necessarily been underneath both the log and the turtles.

For these reasons, these function words are very useful for forensic stylometry. Since the words themselves are so common, a single page of text will have dozens of examples, unlike words like *huge* which may come one to a page if that. People’s characteristic patterns of these words are stable, often over a period of decades (Ryan & Juola, 2011). More importantly, because people pay so little attention to them, they are unlikely to be aware of their own characteristic patterns, and would find it difficult to change their patterns or imitate someone else’s even if so inclined.

However, using this kind of subtle evidence introduces its own issues. It’s simple to say of a word like *chesterfield* that a person either uses it or not. Since literally every English-speaker uses the word *to*, analysts need to dig deeper and ask harder questions, such as how often and in what contexts they use that word, and what other words might be used instead? This opens the door to the kind of statistical analysis described later.

STYLOMETRIC PRACTICE

Using these patterns in practice can be fairly simple. For example, if there are two reasonable suspects, one who uses words with an average length of seven letters and one who uses words with an average length of four, a document that averages 6.95 characters/word is more likely to be by the first author. (A good statistician could put this statement more formally, but the principle is clear.) The main problem with this specific example is that it doesn’t work --- word length isn’t a very stable type of pattern, and a person who uses long words in medical reports is likely to use shorter words in a letter to a friend. More seriously, it doesn’t distinguish well between most people, any more than height does, and for the same reason: most people are of “average” height and “average” vocabulary. It’s also much easier to fake. But measuring many different patterns, it becomes both much harder to fake a dozen or so characteristic patterns at once, and also much more likely that at least a few of those patterns will distinguish usefully between the people of interest.

One of the earliest experiments in this line was the famous study of *The Federalist Papers* by Mosteller and Wallace (1964). These “papers” were a collection of political essays written by influential American politicians in the late 18th century and published pseudonymously under the name “Publius.” The authors, among them James Madison and Alexander Hamilton, later admitted their authorship, but their account of which essays each had written differed. By studying 30 hand-picked function words in undisputed documents, Mosteller and Wallace were able to find significant differences in the use of these words. For example, Madison usually used the word *by* between eleven and thirteen times per thousand words, and never used the word *by* less than five times per thousand. By contrast, Hamilton usually used the word *by* only seven to nine times per thousand words, and never more than thirteen. From this observation alone, it’s easy to infer that a document using *by* only four times per thousand words is by Hamilton, or at least not by Madison. Using a statistical technique called Bayesian analysis, Mosteller and Wallace were able to establish probable authorship of all of the disputed *Federalist Papers*. This landmark paper helped establish statistical text analysis in practice as well as theory.

Another example of this type of analysis is Binongo’s study (2003) of the 15th Oz book. As before, this is a historico-literary and not “forensic” issue, but the issues of a disputed document are similar. The original author of *The Wizard of Oz*, Frank Baum, had died, but the publisher wanted to continue the (profitable) series. The 16th and subsequent books would eventually be written by Ruth Plumly Thompson. But had she also written the 15th book, or had she merely finalized a mostly-complete draft by Baum, as the publisher claimed? Binongo analyzed the fifty most common words in a huge collection of writing by both authors, and applied another statistical method called principle component analysis (PCA) to reduce these fifty variables to a simple visual image (reproduced here as figure 2). This image shows a clear and convincing separation between the works of Baum and of Thompson. Putting the 15th book (*The Royal Book of Oz*) on this same image provides a compelling argument that Thompson, not Baum, penned this work. (Figure 3.)

FIGURE 2 ABOUT HERE

FIGURE 3 ABOUT HERE

Other similar experiments (Holmes, 1998; Holmes and Karsos, 2003; Juola, 2006; Koppel et al., 2009; Stamatatos, 2009; Calle-Martin & Miranda-Garcia, 2012) have confirmed not only the high accuracy of this type of analysis (one well-publicized set of experiments shows better than 90% accuracy at analyzing personal letters) but also that a wide variety of methods and features can be used in stylometry. Examples of such features include not only the use of words, but also the use of phrases, of punctuation, of grammatical patterns and parts of speech, or even of the distribution of individual characters. Similarly, computer scientists have developed many different ways to classify or to categorize data, including language data. Our lab in particular has developed a software program incorporating more than 40,000 useful methods for authorship analysis, most of which can be confirmed to be fairly accurate. (Juola, 2008; Juola, 2012)

One of the best-performing methods we have found is also one of the simplest. As in the previous examples, the analyst first requires a set of documents from all of the prospective authors in the case, just as a fingerprint expert would need sample prints for comparison. He then breaks the documents down to “character n-grams,” groups of n successive characters from among the words in the documents. For example, the word *chairs* would have four 3-grams : *cha*, *hai*, *air*, and *irs*. These character groups would be compiled into a histogram describing the author and his/her most typical character groups. A questioned document is similarly broken down into a histogram describing it, and the author of the most similar histogram is judged to be the most likely author of the disputed document, just as the points that are close together in figure 2 are by the same author. (Noecker & Juola, 2009)

OTHER USES FOR STYLOMETRY

This kind of text analysis can tell us other things as well. Just as individuals have styles, so do groups. A person who leaves his *running shoes* and *toque* near the *chesterfield* is likely to be not only Canadian, but an older Canadian at that (Easson, 2002). Studies (Koppel et al., 2002; Argamon et al., 2009) have shown that similar kinds of analysis to the ones already discussed can distinguish men from women, speakers of different first languages (even when they’re writing in English), different age groups, different education levels, and even different personality types (Gray & Juola, 2009). While a prosecutor may not be interested in the knowledge that the person who wrote a ransom note is most likely a forty year old male native speaker of Spanish, an investigating officer is likely to find that useful, as would a defense attorney tasked with defending a twenty year old female Asian-American against that charge.

NO MAGIC BULLET

Despite this promising set of results, there are still important issues that need to be considered when applying stylistic analysis in any particular situation. Quality issues involving prior investigation, samples, and method should be addressed as early and completely as possible to get maximum use out of any analysis.

Prior Investigation

One of the most important questions is the quality of the prior investigation before the documents even get to the lab bench. The most important question in the minds of many jurists and investigators is “Did this person write that document?” Unfortunately, phrased like that it is almost impossible to answer. While aspects of a person’s writing style can be stable for decades, others can change from one page to the next. Stylometry is much better equipped to answer questions like “which of these people wrote that document?”

A deeper analysis of the nearest-neighbor method of attribution (“the author of the closest unquestioned document is the author of the disputed one”) should make this issue clear. If you have samples from only one undisputed author, then that author will be closest to everything. The analysis will show not only that the defendant wrote the ransom note, but also the Declaration of Independence, *Moby Dick*, and my address book. It is therefore imperative to have a reasonable set of so-called distractor authors, authors who are either other reasonable suspects or at least similar enough to the actual suspect that their writings form a useful and meaningful test.

Another key aspect is the question of “none of the above.” Sometimes one is lucky enough to have the equivalent of a lifeboat murder mystery, one where only a few people had the opportunity to write the document in question, and no stranger could have entered the picture. The case of the dead roommate is likely to be one of those; only a few people would have access to the (locked) room where the computer in question was kept. But often the document comes from a public computer, or from the Internet at large. We might think that so-and-so is a more likely author, but we can’t exclude the possibility of a suspect we didn’t even imagine. Again, the simple nearest-neighbor method (and most other methods) can’t handle this situation; one of the candidate authors will end up being the closest. Handling the none-of-the-above issue is possible (one of my students is just completing a thesis on the subject) but tricky, and care must be taken to avoid muddying the waters.

Sample Quality

Another key issue is the nature of the samples used in the analysis. Of course, everyone knows about the importance of not contaminating a blood sample. But it’s easy to contaminate a writing sample as well. In fact, there are even people who do that kind of thing professionally. They’re called editors. Part of an editor’s job is to “clean up” individual quirks of language, such as spelling errors, malapropisms, and unusual or difficult phrasing. Even typists can unconsciously introduce their own changes, leaving out a word here, inserting a typographical error there, and normalizing the spelling of a word without realizing it. This isn’t a major problem --- obviously it didn’t stop Binongo from studying the highly edited Oz books --- but keep in mind that comparison of edited, published, documents against raw manuscripts can affect the accuracy, and if you have the luxury of using original documents, you should make sure to do so.

Issues of genre need to be considered as well; email is often written in a completely different style to reports, and personal letters fall into yet a third category (Biber, Conrad, and Reppen, 1998). Different genres can confuse an analysis as badly as edited works can. Unfortunately, it’s rare that we have multiple samples of suicide or ransom notes, so direct stylistic comparison can be difficult. While this issue can affect the accuracy of the analysis, good analytic technique can compensate. Again, this is an issue to think about when collecting comparative samples, unless you want to create more work for your analyst.

The third major sample issue is simply quantity. Since much of the work described here is statistical, the computer needs enough writing to have a representative sample. Researchers in this field usually have the luxury of working with recognized and prolific authors (e.g. Baum wrote at least 14 Oz books plus other works, Thompson at least 18) and disputed documents the length of a book. Forensic practitioners often have at most a paragraph. Perhaps obviously, the more data, the better the statistics in the final report will be. Again, this can be compensated to some degree by the analysis method, but investigators should be aware of the issue and try to get as much data as possible.

What method to use?

This type of computer-based analysis has several forensic advantages over simple close reading by language experts. It’s much more objective, and much easier to test. It’s therefore more likely to be able to produce the necessary kind of confidence measures and estimates of error rates that are needed to pass a *Daubert* challenge. The difficulty, ironically enough, is in the specifics of the methods applied. It can be very difficult to justify any one method or choice of parameters, for example using thirty separate features (as in Mosteller-Wallace), fifty (as in Binongo), or even several hundred [as in Hoover (2004a, 2004b)]. Would the results have been substantially different if one used Bayesian analysis instead of nearest-neighbor? Have the methods, in fact, been cherry-picked to deliver the desired results?

We have an easy response to this question as well. With tens of thousands of methods at our disposal and the speed of modern computers, analysis software such as ours can run a thousand analyses as easily as a single one. Using a mixture of different analyses enables the creation of a highly reliable ensemble of analyses. Since different features sets will force the computer to discriminate in different ways, they will possibly produce different answers about the author. But precisely because most of these methods are fairly accurate, they will pick the most common author most of the time.

Here statistics again comes to our rescue. If you flip a weighted coin, one weighted to come down heads 60% of the time, it will land heads most of the time (about 60% of the time, in fact). But if you flip that coin five times, you’ll get at least three heads not 60% of the time, but nearly 70%. If you flip that coin eleven times, you’ll get a majority of heads (six or more) more than 75% of the time. And flipping that coin 51 times will give you a majority of heads (at least 26) more than 90% of the time.

Similarly, if each analysis were only right 60% of the time (a very conservative guess, representing poor data and not much of it), running 51 different and independent analyses would give the right answer as the most common answer more than 90% of the time. A few misanalyses here and there would not affect the overall results. Of course, with more realistic accuracies and better data, the results are even better. With enough methods available, an analyst can afford to run the computer a little longer and be confident of the answer.

SUMMARY

Stylometry isn’t a magic bullet, but it can be a useful technique for gathering evidence where traditional document analysis fails. From its beginnings as a statistical curiosity, it has evolved into a robust and well-studied discipline formalizing what, at base, are simple common sense notions about the idea of how people write and speak differently.

Research, of course, continues. Just as DNA scientists are working to find new ways to analyze smaller and poorer-quality samples, labs across the world are looking for ways to work with poorer and smaller texts and to find more robust markers of individual style. But cases of disputed email are happening today, cases without traditional “documents” for traditional document analysis, and even the current state of the art can develop key evidence in these investigations.

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. OCI-1032683. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge: Cambridge University Press.

Argamon, S., Koppel, M., Pennebaker, J. and Schler, J. (2009). Automatically Profiling the Author of an Anonymous Text. *Communications of the ACM* , 52 (2): 119-123 (virtual extension).

Binongo, J. N. G. (2003). Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance.* 16(2)9—17.

Bransford, J. D., Barlay, J. R., and Franks, J. J. (1972). Sentence memory: A constructive vs. interpretive approach. *Cognitive Psychology.* 3:193–209.

Calle-Martin, J. and Miranda-Garcia, A. (2012). Stylometry and authorship attribution: Introduction to the Special Issue. *English Studies* 93(3):251—58.

Easson, G. (2002). The linguistic implications of shibboleths. *Proc. Annual Meeting of the Canadian Linguistics Association.* Toronto, Canada:CLA

Gray, C. and Juola, P. (2011). Personality identification through textual analysis. *Chicago Colloquium on Digital Humanities and Computer Science (DHCS).* Chicago, IL: University of Chicago.

Holmes, D. and Kardos, J. (2003). Who was the author? An introduction to Stylometry. . *Chance.* 16(2):9—17.

Holmes, D. I. (1998). The evolution of stylometry in humanities computing. *Literary*

*and Linguistic Computing*. 13(3) 111—7.

Hoover, D.L. (2004a). Delta prime?. *Literary and Linguistic Computing*. 19(4): 477—95.

Hoover, D.L. (2004b). Testing Burrows’s Delta. *Literary and Linguistic Computing*. 19(4): 453—75.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology. 60(1):9—26.

Koppel, M., Argamon, S. and Shimoni, A.R. (2002). Automatically categorizing

written texts by author gender. *Literary and Linguistic Computing*. 17(4):401—12.

Juola, P. (2012). Large-scale experiments in authorship attribution. *English Studies* 93(3):275—83.

Juola, P. (2009). 20,000 ways not to do authorship attribution – and a few that work. *Proc. 2009 Biennial Conference of the International Association of Forensic Linguists.* Amsterdam, Netherlands.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*. 1(3):233—334.

Mosteller, F. and Wallace, D. (1964, reprinted 2007). *Inference and Disputed Authorship: The Federalist.* Chicago: University of Chicago Press.

Noecker Jr, J. and Juola, P. (2009). Cosine distance nearest-neighbor classification for authorship attribution. *Proc. Digital Humanities 2009.* College Park, MD: ADHO.

Ryan, M. and Juola, P. (2011). Age or identity: A comparison of stylistic similarity. *Exploring the Boundaries and Applications of Corpus Linguistics*. Tuscaloosa, AL: University of Alabama.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*. 60(4):538—56.

Wellman, F. L. (1936). *The Art of Cross-Examination*. New York:McMillan, 4th edition.

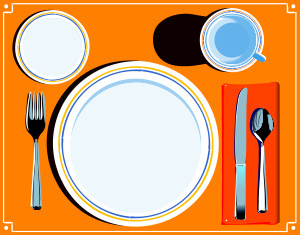
Wikipedia. (n.d.) Athletic shoe. http://en.wikipedia.org/wiki/Athletic\_shoe. Retrieved 9 July 2012.

Patrick Juola, Ph.D., is the Director of Research at Juola & Associates (www.juolaassoc.com), a consulting firm specializing in stylometry, forensic text analysis, and litigation support. He is also the Director of the Evaluating Variation in Language (EVL) laboratory and an associate professor of computer science at Duquesne University, Pittsburgh, PA. He is a Diplomate and Certified Forensic Consultant of the ACFEI.

TEST QUESTIONS:

1. What kind of things can be learned from stylometric analysis?
   1. The author of a document
   2. The age of the author of a document
   3. The education level of the author of a document
   4. The gender of the author of a document
   5. ALL OF THE ABOVE
2. Which of the following did Mosteller and Wallace examine in *the Federalist Papers*?
   1. The fifty most common words
   2. THIRTY HAND-SELECTED FUNCTION WORDS
   3. The average length of sentences
   4. The average length of words
   5. The distribution of character ngrams
3. Which of the following parts of speech are not typically function words?
   1. NOUNS
   2. Pronouns
   3. Prepositions
   4. Conjunctions
   5. All of the above are usually function words
4. Function words are particularly informative because
   1. PEOPLE DON’T REALLY NOTICE THEM AND THUS THEY’RE HARD TO FAKE
   2. They’re rare and thus particularly informative
   3. Only educated people use them
   4. They’re long words that offer opportunities for mistakes
   5. They begin with capital letters and are easy to spot
5. Which of the following does not normally impact the quality of samples for a stylometric analysis?
   1. The amount of data available for study
   2. Whether or not a document has been edited
   3. The type or genre of a document
   4. WHETHER SOMEONE OTHER THAN THE NAMED SUSPECTS MIGHT HAVE WRITTEN THE QUESTIONED DOCUMENT
   5. Nothing impacts the quality of samples; they’re always perfect
6. Stylometric text analysis …
   1. … should be a magic bullet for all issues of disputed documents
   2. … is an arcane statistical nightmare that confuses more than it explains
   3. … CAN BE USEFUL FOR ANSWERING QUESTIONS WHERE TRADITIONAL DOCUMENT ANALYSIS FAILS
   4. … was invented in the 1990s
   5. … is unlikely to pass a *Daubert* challenge because it has not been subject to peer review and publication

Figure 1: Public domain image courtesy of www.wpclipart.com



A place setting : where is the fork?

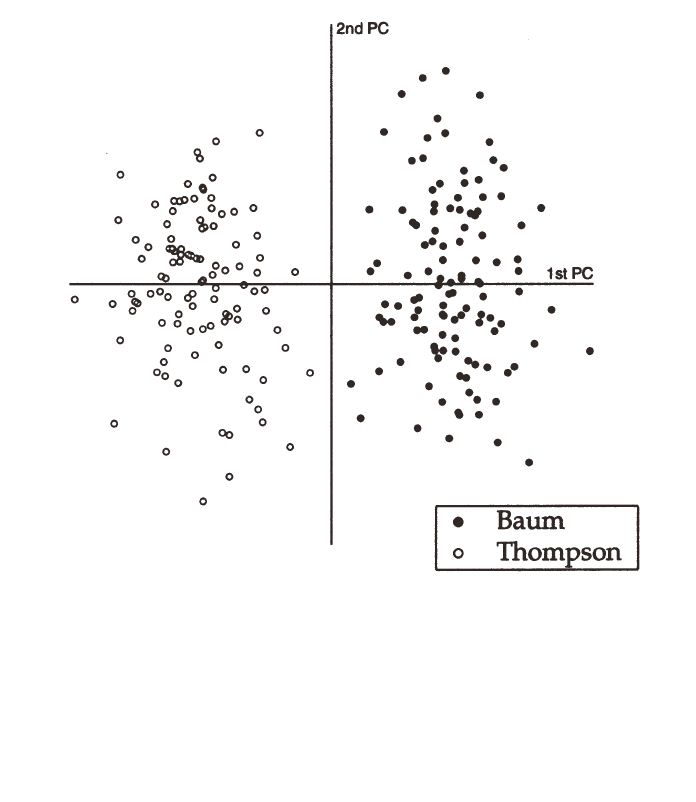
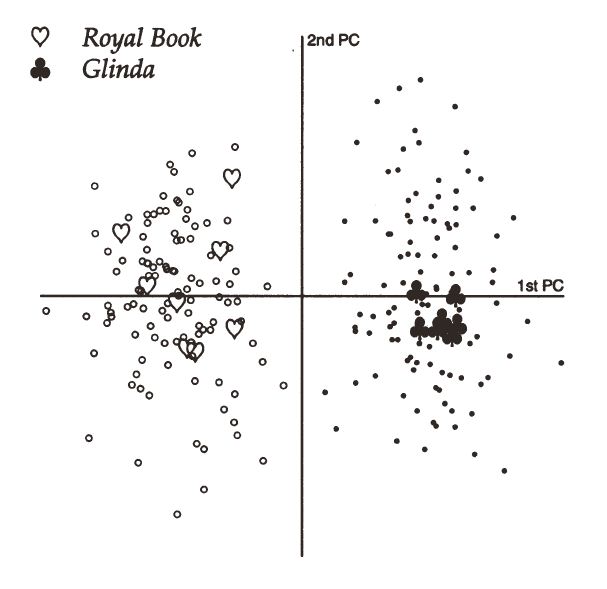
Figure 2: From (Binongo, 2003), used by permissionPCA analysis of writings by Baum and Thompson. Note clear and convincing visual grouping and separation.

Figure 3: From (Binongo, 2003), used by permission



Analysis as in figure 1 of two disputed works (*Royal Book of Oz* and *Glinda of Oz*). The inference is that *Royal Book* is by Thompson, *Glinda* by Baum.