

Summary: The Effectiveness of the Stylometry of Function Words in
Discriminating between Shakespeare and Fletcher

Thomas Bolton Horton
University of Edinburgh, Department of Computer Science
Ph.D. Dissertation, 1987

A number of recent successful authorship studies have relied on a statistical analysis of language features based on function words. However, stylometry has not been extensively applied to Elizabethan and Jacobean dramatic questions. To determine the effectiveness of such an approach in this field, language features were studied in twenty-four plays by Shakespeare and eight by Fletcher. The goal was to develop procedures that might be used to determine the authorship of individual scenes in *The Two Noble Kinsmen* (TNK) and *Henry VIII* (H8).

Of the 32 texts of known authorship, 6 were set aside as a test set. These were treated as if their authorship was unknown. All procedures that I evaluated were applied to samples from these 6 plays, and these results were used to judge their effectiveness. The remaining 20 Shakespeare texts and the 6 Fletcher texts made up the control set, which was used to establish each dramatist's characteristics of composition.

Homonyms, spelling variants and contracted forms in old-spelling dramatic texts present problems for a computer analysis. Many common function words have several variant spellings (e.g. "been" can be spelled "beene", "bene", "bin" etc). Other forms can represent a number of lexical forms; for example, besides the indefinite article, the single-letter word "a" can mean "he," "of," "on," "ah" etc. Some forms of compound contractions involving function words are frequent (for example, "let's", "o'th'", "'tis" and the many contracted forms of "is" like "it's", "Caesar's", "he's" etc).

A program (called REPLACE) that uses a system of pre-edit codes and replacement/expansion lists was developed to prepare versions of the texts in which all forms of common words can be recognized automatically. Homonyms and variants were found and marked in each text using a system of hash suffixes. For example, "a#1" represents occurrences of "he", "a#3" represents "on", etc. Occurrences of "beene", "bene", etc are replaced by the standard form (but occurrences of "bin" meaning a container are marked "bin#1"). In addition, a list of compound contractions and their full forms was compiled (from experience and with help from Partridge's book *Orthography in Shakespeare and Elizabethan Drama*). A simple replacement strategy is not powerful enough to handle apostrophe-s and -t forms. Hash suffixes were also used to distinguish apostrophe-s contractions involving "is", "us", "his" etc from possessive forms, and contractions of "it" from forms like "banish't" (ie "banished").

Program REPLACE made use of these special markings and the expansion lists to prepare "expanded" versions of the plays. These versions were then used to determine the extent of each author's use of compound contractions. In almost every case, Fletcher uses more of these forms than Shakespeare, although the latter uses more contractions as his career progresses. Because of this secular change and the possibility of alterations introduced by scribes, compositors or revisors, the expanded versions of the plays were used in remainder of the study.

To evaluate some procedures for determining authorship developed by A. Q. Morton and his colleagues, occurrences of 30 common collocations and 5 proportional pairs are analyzed in the texts. Within-author variation for these features is greater than had been found in previous studies. Univariate chi-square tests are shown to be of limited usefulness because of the statistical distribution of these textual features and correlation between pairs of features. Contrary to some earlier claims, the best of the collocations do not discriminate as well as most of the individual words from which they are composed.

Turning to the rate of occurrence of individual words and groups of words, distinctiveness ratios and t-tests are used to select variables that best discriminate between Shakespeare and Fletcher. Variation due to date of composition and genre within the Shakespeare texts is examined using the statistical procedure analysis of variance.

A number of potential markers of authorship were eliminated because the rate of use for one of the subgroups (such as comedies, or late plays) was too close to Fletcher's overall rate. Some of the observed variations are interesting in their own right. Shakespeare's comedies are characterized by high rates for pronouns and "a", together with low rates for "the". Tragedies have low rates for "a". The histories have very low rates for personal pronouns (as noted by Brainerd) and high rates for "in", "of" and "and". "In" occurs infrequently in the romances, while "so" is much more frequent in this genre than the other three. The late plays have a high rate for "the".

The rates for several word classes were examined (pronouns, forms of "have", "be" and "do", and modal verbs), but none of these group variables proved useful. However, when compiling the list of spelling variants, I noticed that Shakespeare uses more forms that begin with "there-" or "where-" (such as "therefore", "therein", "wheresoever", etc); the forms that Fletcher does use occur less frequently than in Shakespeare's texts. When the rates for all these "there/where- compounds" are combined, Shakespeare's overall rate is almost 12 times that of Fletcher. These forms are rather infrequent, and some forms appear in stock phrases or songs and may not reflect the author's normal usage; for these reasons, this group was not used as a variable in the main analysis of function words. However, two scenes usually attributed to Fletcher, *Henry VIII* I.iii and *The Two Noble Kinsmen* IV.iii, contain what I feel are significant occurrences of there/where-compounds. (Later analysis indicated that the use of function words in both scenes is also much closer to Shakespeare's known work.)

A multivariate and distribution-free discriminant analysis procedure (using kernel estimation) was used to determine if data from a single scene resembled the scenes from Shakespeare or Fletcher more closely. The classifiers based on the best marker words and the kernel method were carefully evaluated using the texts of known authorship. To study the effect of characterization, I extracted the speeches of 62 characters (who speak at least 500 words) from 6 test-set plays. The procedure was only slightly less accurate in classifying these character samples than the set of test-set scenes, which suggests that characterization does not affect the use of these word-rate variables with this procedure to any great degree (at least for the purpose of distinguishing Fletcher from Shakespeare). I also tested the procedures with smaller and smaller scenes, and found that they performed well for samples as short as 500 words. When the final procedure is used to assign the 459 scenes of known authorship (containing at least 500 words), 94.8% are assigned to the correct author. Only two scenes are incorrectly classified, and 4.8% of the scenes cannot be assigned to either author by the procedure.

When applied to individual scenes of at least 500 words in *The Two Noble Kinsmen* and *Henry VIII*, the procedure indicates that both plays are collaborations and generally supports the usual division. However, the marker words in a number of scenes often attributed to Fletcher are very much closer to Shakespeare's pattern of use. Some of the more interesting results include the assignment to Shakespeare of TNK IV.iii, which most scholars have regarded as decent imitation. (The function word result for this scene is supported to some extent by an occurrence of a there/where- compound and by rates for "hath" and "you" that are unlike rates in known Fletcher scenes.) In H8, I.iii is another scene that resembles Shakespeare in the use of function words and there/where-compounds. It also contains a number of convincing Fletcher stylistic traits, so perhaps revision should be proposed.

The function-word results for Shakespeare are extremely strong for the prose scene V.iv, which contains occurrences of "ye" and "em" (Fletcher traits) that are unparalleled in Shakespeare's texts. This contradictory evidence raises questions about the copy-text; again, revision is a plausible explanation (although the results are so strong in this case that the scene appears to be mainly Shakespeare's work). The function-word results support Foakes' and Hoy's suggestions that Fletcher touched up Shakespeare's work in II.i-ii, III.iib, and IV.i-ii. The use of function words in Act IV is very unlike Fletcher, and my results indicate that he had little or nothing to do with it. Results for II.ii and III.iib are less clear but may support the theory of revision.

The contents of this electronic file are copyright ©1990 Thomas B. Horton. Quotation for scholarly (non-commercial) purposes is permitted, but please contact the author to verify the material in question and advise him of your intention. Distribution is only permitted if the file is not changed in any way and if this notice is included in the file. The author may be contacted by e-mail at [\[redacted\]](#) or by regular mail at: Department of Computer Science, Florida Atlantic University, Boca Raton, FL 33431 USA