Automatic Compositor Attribution in the First Folio of Shakespeare

Maria Ryskina* **Hannah Alpert-Abrams**† **Dan Garrette**‡ **Taylor Berg-Kirkpatrick***

* Language Technologies Institute, Carnegie Mellon University, {mryskina,tberg}@cs.cmu.edu
†Comparative Literature Program, University of Texas at Austin, halperta@gmail.com
‡ Google, dhgarrette@google.com

Abstract

Compositor attribution, the clustering of pages in a historical printed document by the individual who set the type, is a bibliographic task that relies on analysis of orthographic variation and inspection of visual details of the printed page. In this paper, we introduce a novel unsupervised model that jointly describes the textual and visual features needed to distinguish compositors. Applied to images of Shakespeare's *First Folio*, our model predicts attributions that agree with the manual judgements of bibliographers with an accuracy of 87%, even on text that is the output of OCR.

1 Introduction

Within literary studies, the field of bibliography has an unusually long tradition of quantitative analysis. One particularly relevant area is that of *compositor attribution*—the clustering of pages in a historical printed document by the individual (the *compositor*) who set the type. Like stylometry, a long-standing area of NLP that has largely focused on attributing the authorship of text (Holmes, 1994; Hope, 1994; Juola, 2006; Koppel et al., 2009; Jockers and Witten, 2010), the analysis of *orthographic* patterns is fundamental to compositor attribution. Additionally, compositor attribution often makes use of *visual* features, such as whitespace layout, introducing new challenges. These analyses have traditionally been done by hand, but efforts are painstaking due to the difficulty of manually recording these features.

In this paper, we present an unsupervised model specifically designed for compositor attribution that incorporates both textual and visual sources of evidence traditionally used by bibliographers (Hinman, 1963; Taylor, 1981; Blayney, 1991).

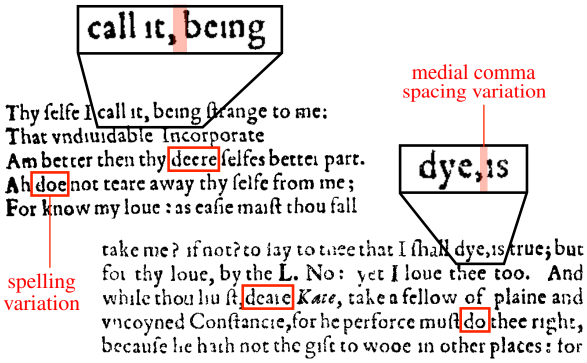
call ıt, beıng
medial comma
spacing variation
Thy ſelfe I call ıt, beıng ſtrange to me:
That vndiuidable Incorporate
Am better then thy deere ſelfes better part.
Ah doe not teare away thy ſelfe from me;
For know my loue : as eaſie maiſt thou fall
dye, ıs
spelling
variation
take me? ıf not? to ſay to thee that I ſhall dye, ıs true; but
for thy loue, by the L. No: yet I loue thee too. And
while thou liu ſt, deare *Kate*, take a fellow of plaine and
vncoyned Conſtancie, for he perforce muſt do thee right,
becauſe he hath not the gift to wooe ın other places: for

Figure 1: The compositor of the left page tended to use the spellings doe and deere, while the compositor for the right page used spellings do and deare, indicating these pages were likely set by different people. The varying width of the medial comma whitespace also distinguishes the typesetters.

Our model jointly describes the patterns of variation both in orthography and in the whitespace between glyphs, allowing us to cluster pages by discovering patterns of similarity and difference. When applied to digital scans of historical printed documents, our approach learns orthographic and whitespace preferences of individual compositors and predicts groupings of pages set by the same compositor.[1] This is, to our knowledge, the first attempt to perform compositor attribution automatically. Prior work has proposed automatic approaches to authorship attribution—which is typically viewed as the supervised problem of identifying a particular author given samples of their writing. In contrast, compositor attribution lacks supervision because compositors are unknown and, in addition, focuses on different linguistic patterns. We explain spellings of words conditioned on word choice, not the word choice itself.

[1] The validity of compositor attribution has sparked an ongoing and heated debate among bibliographers (McKenzie, 1969, 1984; Rizvi, 2016); while some reject parts or all of this approach, it continues to be cited in authoritative bibliographical texts (Gaskell, 2007; Blayney, 1996). Without taking a position in this debate, we seek only to automate the methods that remain in use by particular bibliographers (Blayney, 1996; Burrows, 2013).

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 411–416
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-2065

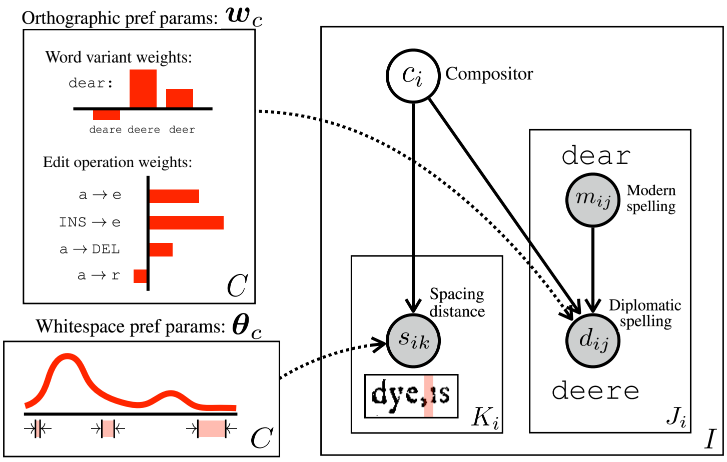

Figure 2: In our model, a compositor c_i is generated for page i from a multinomial prior. Then, each diplomatic word, d_{ij}, is generated conditioned on c_i and the corresponding modern word, m_{ij}, from a distribution parameterized by weight vector $\boldsymbol{w}_c$. Finally, each medial comma spacing width (measured in pixels), s_{ik}, is generated conditioned on c_i from a distribution parameterized by $\boldsymbol{\theta}_c$.

To evaluate our approach, we fit our model to digital scans of Shakespeare's *First Folio* (1623)—a document with well established manual judgements of compositor attribution. We find that even when relying on noisy OCR transcriptions of textual content, our model predicts compositor attributions that agree with manual annotations 87% of the time, outperforming several simpler baselines. Our approach opens new possibilities for considering patterns across a larger vocabulary of words and at a higher visual resolution than has been possible historically. Such a tool may enable scalable first-pass analysis in understudied domains as a complement to humanistic studies of composition.

2 Background

In this paper we focus on modeling the same types of observations made by scholars and demonstrate agreement with authoritative attributions. We use compositor studies of Shakespeare's *First Folio* to inform our approach, drawing on the methods proposed by Hinman (1963), Howard-Hill (1973), and Taylor (1981). Hinman's landmark 1963 study clustered the pages of the *First Folio* according to five different compositors based on variations in spelling among three common words. Figure 1, for example, shows portions of two pages of the *First Folio* with different spelling variants for the words `dear` and `do`: one compositor used `deere` and `doe`, while the other used `deare` and `do`. Hinman relied on the assumption that each compositor was consistent in their preferences for the sake of convenience in the typesetting process (Blayney, 1991). Subsequent studies looked at larger sets of words and more general orthographic preferences (e.g. the preference to terminate words with `-ie` instead of `-y`), leading to modifications of Hinman's original analysis (Howard-Hill, 1973; Taylor, 1981). In this paper we propose a probabilistic model designed to capture both word-specific preferences and general orthographic patterns. To separate the effect of the compositor from the choices made by the author or editor, we condition on a modernized (collated) version of Shakespeare's text as was done by scholars.

Visual features, including typeface usage and whitespace layout, also inform compositor attribution. For example, the highlighted spacing in Figure 1 shows different choices after medial commas (commas that occur before the end of the line). Bibliographers produced new hypotheses about how many compositors were involved in production based on the analysis of the use of spaces before and after punctuation (Howard-Hill, 1973; Taylor, 1981). We additionally incorporate this source of evidence into to our automatic approach by modeling pixel-level whitespace distances.

Bibliographers also use contextual information to inform their analyses, including copy text orthography, printing house records, collation, type case usage, and the use of type with cast-on spaces. In our model, we restrict our analysis to only those features that can be derived from the OCR output and simple visual analysis.

3 Model

Our computational approach to compositor attribution operates on the sources of evidence that have been considered by bibliographers. In particular, we focus on jointly modeling patterns of orthographic variation and spacing preferences across pages of a document, treating compositor assignments as latent variables in a generative model. We assume access to a diplomatic transcription of the document (a transcription faithful to the original orthography), which we automatically align with a modernized version.[2] We experiment with both manually and automatically (OCR) produced transcriptions, and assume access to pixel-level spacing information on each page, which can be extracted using OCR as described in Section 4.

Figure 2 shows the generative process. In our model, each of I total pages is generated independently. The compositor assignment for the ith page is represented by the variable $c_i \in \{1, \ldots, C\}$ and is generated from a multinomial prior. For page i, each diplomatic word, d_{ij}, is generated conditioned on the corresponding modern word, m_{ij}, and the compositor who set the page, c_i. Finally, the model produces the pixel width of the space after each medial comma, s_{ik}, again conditioned on the compositor, c_i. The joint distribution for page i, conditioned on modern text, takes the following form:

$$
\begin{aligned}
&P(\{d_{ij}\}, \{s_{ik}\}, c_i | \{m_{ij}\}) = \\
&\quad P(c_i) && \text{[Prior on compositors]} \\
&\quad \cdot \prod_{j=1}^{J_i} P(d_{ij} | m_{ij}, c_i; \boldsymbol{w}_{c_i}) && \text{[Orthographic model]} \\
&\quad \cdot \prod_{k=1}^{K_i} P(s_{ik} | c_i; \boldsymbol{\theta}_{c_i}) && \text{[Whitespace model]}
\end{aligned}
$$

3.1 Orthographic Preference Model

We choose the parameterization of the distribution of diplomatic words in order to capture two types of spelling preference: (1) general preferences for certain character groups (such as `-ie`) and (2) preferences that only pertain to a particular word and do not indicate a larger pattern. Since it is unknown which of the two behaviors is dominant, we let the model describe both and learn to separate their effects. Using a log-linear parameterization,

$$P(d|m, c; \boldsymbol{w}) \propto \exp(\boldsymbol{w}_c^\top \mathbf{f}(m, d))$$

we introduce features to capture both effects. Here, $\mathbf{f}(m, d)$ is a feature function defined on modern word m paired with diplomatic word d, while $\boldsymbol{w}_c$ is a weight vector corresponding to compositor c.

To capture word-specific preferences we add an indicator feature for each pair of modern word m and diplomatic spelling d. We refer to these as WORD features below. To capture general orthographic preferences we introduce an additional set of features based on the edit operations involved in the computation of Levenshtein distance between m and d. In particular, each operation is added as a separate feature, both with and without local context (previous or next character of the modern word). We refer to this group as EDIT features. The weight vector for each compositor represents their unique biases, as shown in the depiction of these parameters in Figure 2.

3.2 Whitespace Preference Model

Manual analysis of spacing has revealed differences across pages. In particular, the choice of spaced or non-spaced punctuation marks is hypothesized by biobliographers to be indicative of compositor preference and specific typecase. We add whitespace distance to our model to capture those observations. While bibliographers only made a coarse distinction between spaced or non-spaced commas, in our model we generate medial comma spacing widths, s_{ik}, that are measured in pixels to enable finer-grained analysis. We use a simple multinomial parameterization where each pixel width is treated as a separate outcome up to some maximum allowable width:

$$s_{ik} | c_i \sim Mult(\boldsymbol{\theta}_{\boldsymbol{c_i}})$$

Here, $\boldsymbol{\theta}_c$ represents the vector of multinomial spacing parameters corresponding to compositor c. We choose this parameterization because it can capture non-unimodal whitespace preference distributions, as depicted in Figure 2, and it makes learning simple.

3.3 Learning and Inference

Modern and diplomatic words and spacing variables are observed, while compositor assignments are latent. In order to fit the model to an input document we estimate the orthographic preference parameters, $\boldsymbol{w}_c$, and spacing preference parameters, $\boldsymbol{\theta}_c$, for each compositor using EM. The E-step is accomplished via a tractable sum over compositor assignments, while the M-step for $\boldsymbol{w}_c$ is accomplished via gradient ascent (Berg-Kirkpatrick et al., 2010). The M-step for spacing parameters, $\boldsymbol{\theta}_c$, uses the standard multinomial update. Predicting compositor groups is accomplished via an

[2] Modern editions are common for many books that are of interest to bibliographers, though future work could consider how to cope with their absence.

Model Setup	Bodleian Transcription				Ocular OCR Transcription			
	Hinman Attr		Blayney Attr		Hinman Attr		Blayney Attr	
	1-to-1	M-to-1	1-to-1	M-to-1	1-to-1	M-to-1	1-to-1	M-to-1
RANDOM	22.5	49.6	16.7	49.6	22.5	49.6	16.7	49.6
BASIC w/ HINMAN	67.9	71.8	60.4	67.3	66.6	70.5	47.1	63.8
w/ AUTO	64.3	81.0	58.8	81.3	64.9	81.1	53.7	80.7
FEAT w/ EDIT	75.3	79.1	77.1	83.1	76.8	77.4	76.1	76.0
w/ EDIT + WORD	81.1	81.1	80.7	80.6	75.1	75.0	74.4	74.4
w/ EDIT + SPACE	**87.6**	**87.5**	**87.3**	**87.2**	**86.7**	**86.6**	**85.9**	**85.8**
w/ ALL	83.8	83.7	83.5	83.4	82.5	82.4	82.4	82.2

Table 1: The experimental results for all setups of the model. In the experiments with BASIC model, we compare the short HINMAN word list with the automatically filtered AUTO word list. We show results for several variants of our full model, labeled as FEAT, both with and without spacing generation. A random baseline is included for comparison.

independent argmax over each c_i. In all experiments we run 75 iterations of EM with 100 random restarts, choosing the learned parameters corresponding the best model likelihood.

4 Experiments

Data: To evaluate our model when it has access to perfectly transcribed historical text, we use the Bodleian diplomatic transcription of the *First Folio*.[3] To test whether our approach can also work with untranscribed books, we ran the Ocular OCR system (Berg-Kirkpatrick et al., 2013) on the Bodleian facsimile images to create an automatic diplomatic transcription. In both cases, we used Ocular's estimates of glyph bounding boxes on the complete *First Folio* images to extract spacing information. The modern text was taken from MIT Complete Works of Shakespeare[4] and was aligned with diplomatic transcriptions by running a word-level edit distance calculation. The extracted substitutions form the model's observed modern-diplomatic word pairs.

Evaluation: To compare the recovered attribution with those proposed by bibliographers, we evaluate against an authoritative attribution compiled by Peter Blayney (1996) which includes the work of various scholars (Hinman, 1963; Howard-Hill, 1973, 1976, 1980; Taylor, 1981; O'Connor, 1975; Werstine, 1982). We also evaluate our system against an earlier, highly influential model proposed by Hinman (1963), which we approximate by reverting certain compositor divisions in Blayney's attribution. Hinman's attribution posited five compositors, while Blayney's posited eight. In experiments, we set the model's maximum number of compositors to $C = 5$ when evaluating on Hinman's attribution, and use $C = 8$ with Blayney's. We compute the one-to-one and many-to-one accuracy, mapping the recovered page groups to the gold compositors to maximize accuracy, as is standard for many unsupervised clustering tasks, e.g. POS induction (see Christodoulopoulos et al. (2010)).

BASIC model variant: We evaluate a simple baseline model that uses a multinomial parameterization for generating diplomatic words and does not incorporate spacing information. We use two different options for selection of spelling variants to be considered by the model. First, we consider only the three words selected by Hinman: `do`, `go` and `here` (referred to as HINMAN). Second, we use a larger, automatically selected, word list (referred to as AUTO). Here, we select all modern words with frequency greater than 70 that are not names and that exhibit sufficient variance in diplomatic spellings (most common diplomatic spelling occurs in less than 80% of aligned tokens). For our full model, described in the next section, we always use the larger AUTO word list.

FEAT model variant: We run experiments with several variants of our full model, described in Section 3 (referred to as FEAT since they use a feature-based parameterization of diplomatic word generation.) We try ablations of WORD and EDIT features, as well as model variants with and without the spacing generation component (referred to as SPACE.) We refer to the full model that includes both types of features and spacing generation as ALL.

5 Results

Our experimental results are presented in Table 1. The BASIC variant, modeled after Hinman's original procedure, substantially outperforms the random baseline, with the HINMAN word list outperforming the larger AUTO word list. However, use

[3] http://firstfolio.bodleian.ox.ac.uk/

[4] http://shakespeare.mit.edu/

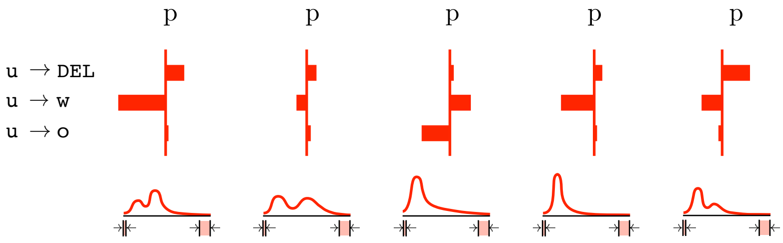

Figure 3: Learned behaviors of the Folio compositors. Our model only detected the presence of five compositors (ranked according to number of pages the compositor set in our model's prediction). Compositor D's habit of omitting `u` (`yong` vs. `young`) and compositor C's usage of spaced medial commas were also noticed in Taylor (1981).

of the larger word list with feature-based models yields large gains in all scenarios, including evaluation on Hinman's original attributions and while using OCR diplomatic transcriptions. The best-performing model for both manually transcribed and OCR text uses EDIT features in conjunction with spacing generation and achieves an accuracy of up to 87%. Including WORD features on top of this leads to slightly reduced performance, perhaps as a result of the substantially increased number of free parameters. In the OCR scenario, the addition of WORD features on top of EDIT decreases accuracy, unlike the same experiment with the manual transcription. This is possibly a result of the reduced reliability of full word forms due to mistakes in OCR.

Particularly interesting is the result that spacing, rarely a factor considered in NLP models, improves the accuracy significantly for our system when compared with EDIT features alone. Because pixel-level visual information and arbitrary orthographic patterns are also the most difficult features to measure manually, our results give strong evidence to the assertion that NLP-style models can aid bibliographers.

6 Discussion

The results on OCR (character error rate for most plays $\approx 10-15\%$) transcripts are only marginally worse than those on manual transcripts, which shows that our approach can be generalized for the common case where manual diplomatic transcriptions are not available. For our experiments, we also chose a common modern edition of Shakespeare instead of more carefully produced modernized transcription of the facsimile—our goal being to again show that this approach can be generalized, perhaps to documents where careful modernizations of the facsimile are not available. Together, these results suggest that our model may be sufficiently robust to aid bibliographers in their analysis of less studied texts.

Figure 3 shows an example of the feature weights and spacing parameters learned by the FEAT W/ ALL model. Our statistical approach is able to successfully explain some of the observations scholars made. For example, Taylor (1981) notices that compositors C and D prefer to omit `u` in `young` but A does not. Our model reflects this by giving `u` → `DEL` high weight for D and low weight for A. However, the weight of a single feature is difficult to interpret in isolation. This might be the reason why our model only moderately agrees in case of compositor C. Another example can be seen in spacing patterns: according to Taylor (1981), compositor C uses spaced medial commas unlike A and D. Our model learns the same behavior.

7 Conclusion

Our primary goal is to scale the methods of compositor attribution, including both textual and visual modes of evidence, for use across books and corpora. By using principled statistical techniques and considering evidence at a larger scale, we offer a more robust approach to compositor identification than has previously been possible. The fact that our system works well on OCR texts means that we are not restricted to only those documents for which we have manually produced transcriptions, opening up the possibility for bibliographic study on a much larger class of texts. Though we are unable to incorporate the kinds of world knowledge used by bibliographers, our ability to include more information and more fine-grained information allows us to recreate their results. Having validated these techniques on the *First Folio*, where historical claims are well established, we hope future work can extend these methods and their application.

Acknowledgements

We thank the three anonymous reviewers for their valuable feedback. This project is funded in part by the NSF under grant 1618044.

References

Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of the Association for Computational Linguistics*.

Peter W. M. Blayney. 1991. *The First Folio of Shakespeare: In Conjunction with the Exhibition at the Folger Shakespeare Library, Washington, DC, April 1, 1991-September 21, 1991*. Folger Shakespeare Library.

Peter W. M. Blayney, editor. 1996. *The First Folio of Shakespeare: The Norton Facsimile*. Norton.

Ian R. Burrows. 2013. "The peryod of my blisse": Commas, ends and utterance in Solyman and Perseda. *Textual Cultures: Texts, Contexts, Interpretation* 8(2):95–120.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. 2010. Two decades of unsupervised POS induction: How far have we come? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Philip Gaskell. 2007. *A New Introduction to Bibliography*. Oak Knoll Press.

Charlton Hinman. 1963. *The printing and proofreading of the First Folio of Shakespeare*, volume 1. Oxford: Clarendon Press.

David I. Holmes. 1994. Authorship attribution. *Computers and the Humanities* 28(2):87–106.

Jonathan Hope. 1994. *The authorship of Shakespeare's plays: A socio-linguistic study*. Cambridge University Press.

Trevor H. Howard-Hill. 1973. The compositors of Shakespeare's Folio Comedies. *Studies in bibliography* 26:61–106.

Trevor H. Howard-Hill. 1976. *Compositors B and E in the Shakespeare First Folio and Some Recent Studies*. Self-published.

Trevor H. Howard-Hill. 1980. New light on compositor E of the Shakespeare First Folio. *The Library* 6(2):156–178.

Matthew L. Jockers and Daniela M. Witten. 2010. A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing* 25(2):215–223.

Patrick Juola. 2006. Authorship attribution. *Foundations and Trends in Information Retrieval* 1(3):233–334.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Technol.* 60(1):9–26. https://doi.org/10.1002/asi.v60:1.

D. F. McKenzie. 1969. Printers of the mind: Some notes on bibliographical theories and printing-house practices. *Studies in Bibliography* 22:1–75.

D. F. McKenzie. 1984. Stretching a point: Or, the case of the spaced-out comps. *Studies in Bibliography* 37:106–121.

John O'Connor. 1975. Compositors D and F of the Shakespeare First Folio. *Studies in Bibliography* 28:81–117.

Pervez Rizvi. 2016. Use of spellings for compositor attribution in the First Folio. *The Papers of the Bibliographical Society of America* 110:1–53.

Gary Taylor. 1981. The shrinking compositor A of the Shakespeare First Folio. *Studies in Bibliography* 34:96–117.

Paul Werstine. 1982. Cases and compositors in the Shakespeare First Folio Comedies. *Studies in Bibliography* 35:206–234.