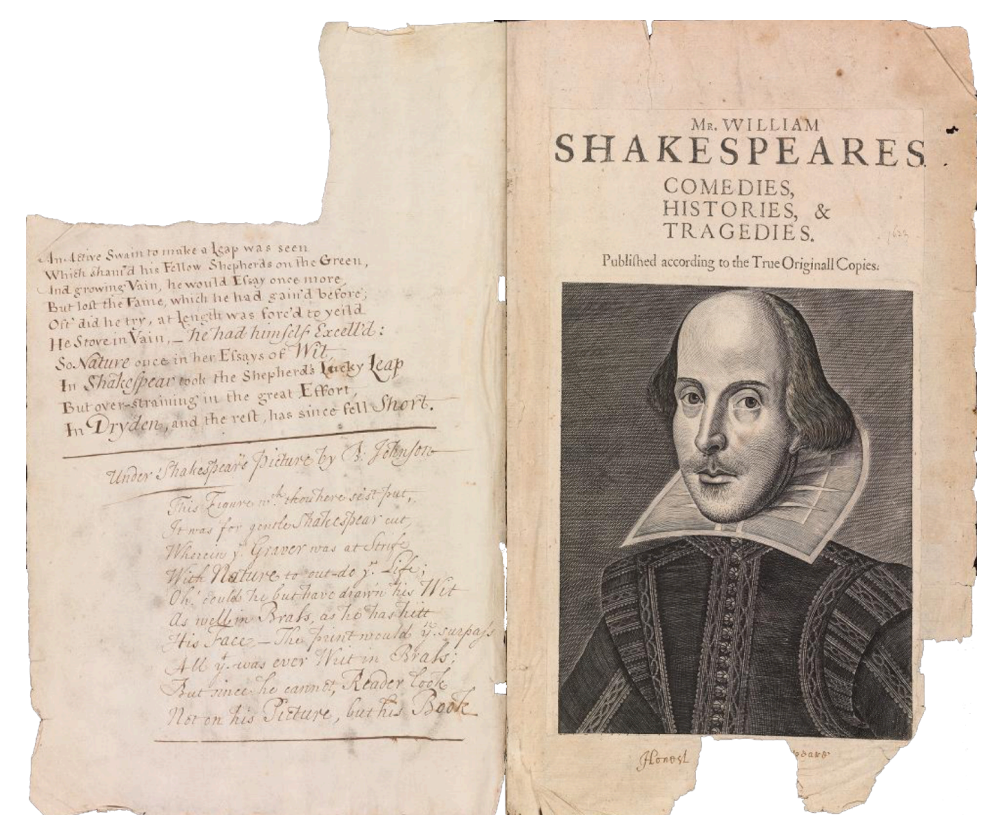
Automatic Compositor Attribution in the First Folio of Shakespeare

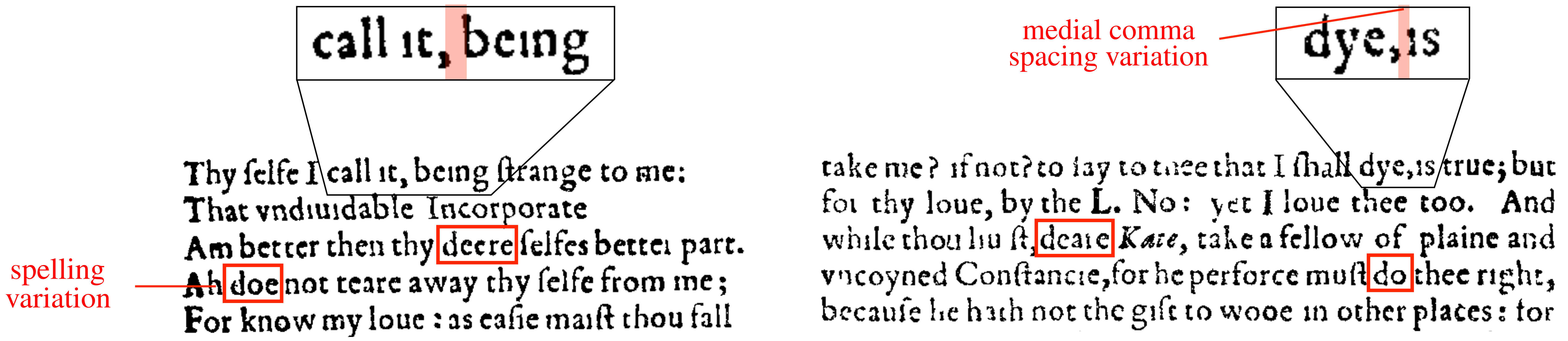
Maria Ryskina*, Hannah Alpert-Abrams†, Dan Garrette‡ and Taylor Berg-Kirkpatrick*
*{mryskina, tberg}@cs.cmu.edu †halperta@gmail.com ‡dhgarrette@google.com

Summary



We present an unsupervised approach to *compositor attribution* — clustering the pages of a historical printed document according to the individual who set the type. We use the *First Folio of Shakespeare* (1623) as a test case since it has been extensively studied by bibliographers. Following the manual work of these traditional Shakespeare scholars, we perform automatic analysis by modeling the orthographic preferences and spacing tendencies of compositors, reaching up to 87% agreement with the authoritative attribution.

What is a compositor?

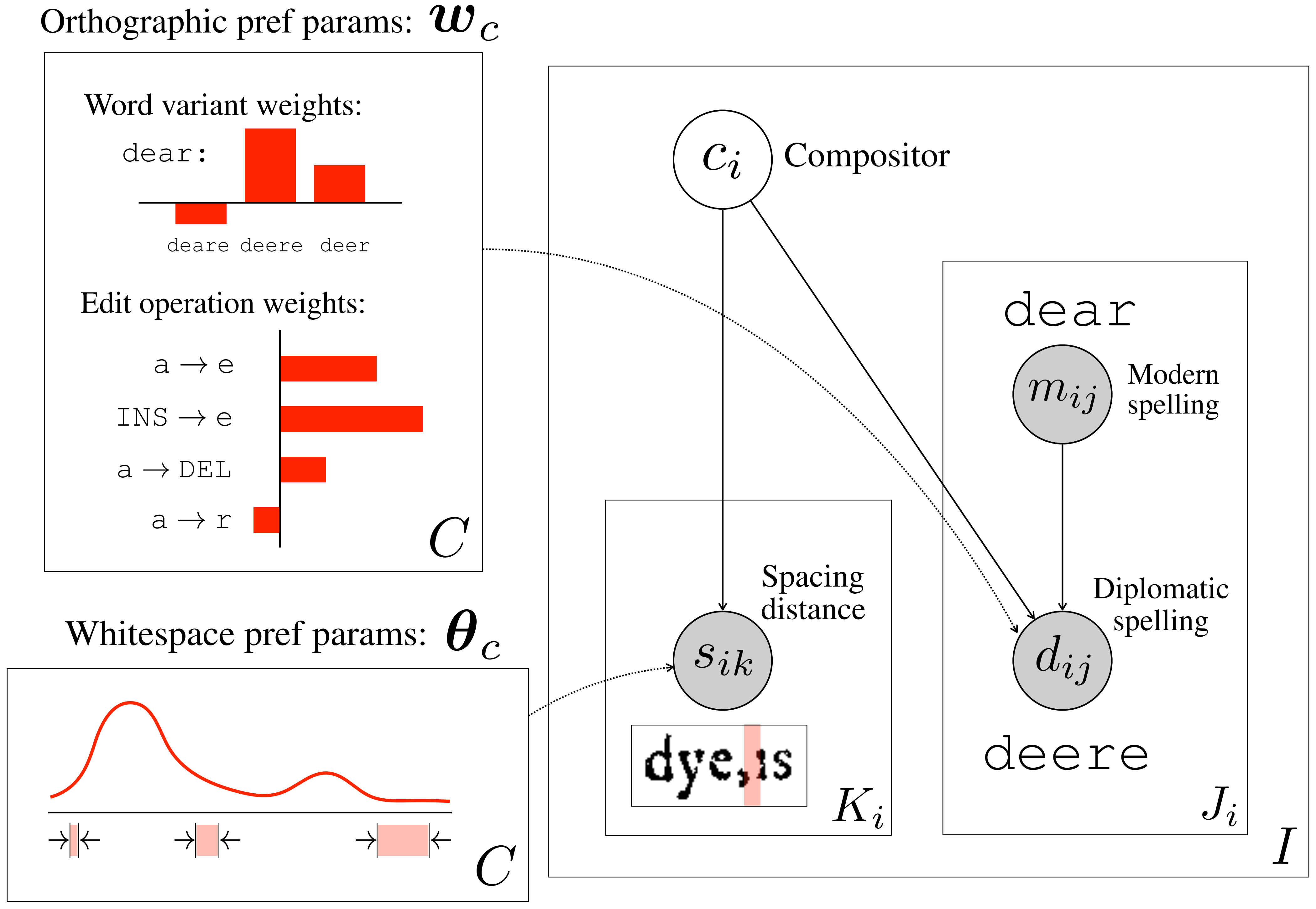

call it, being
dye,is
medial comma
spacing variation
spelling
variation

Thy ſelfe I call it, being 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

take me? if not? to ſay to thee that I ſhall dye,is 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 in other places: for

A compositor is a person who manually arranges and sets type for printing a document. Shakespeare's First Folio is believed to have been set by multiple compositors, each with varying degrees of proficiency at accurately transcribing the original (mostly lost) manuscripts. Bibliographers attribute pages to compositors based on their spelling choices or visual evidence such as whitespace lengths before and after punctuation.

Model

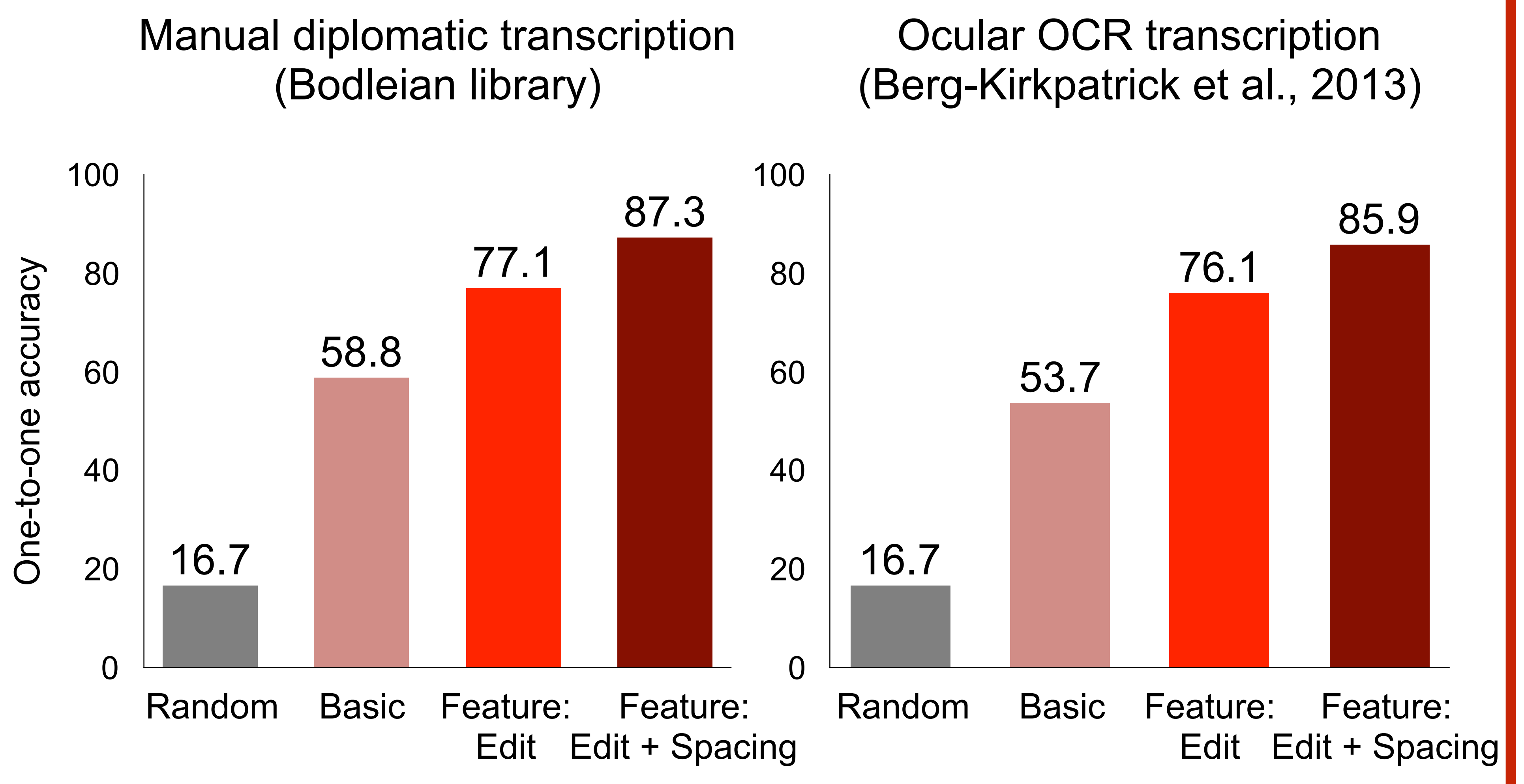


Compositors are modeled as latent variables, one per page, and their orthographic and spacing choices are modeled by multinomial distributions.

Basic model variant:
Simple multinomial baseline, only includes word-level spelling choices.

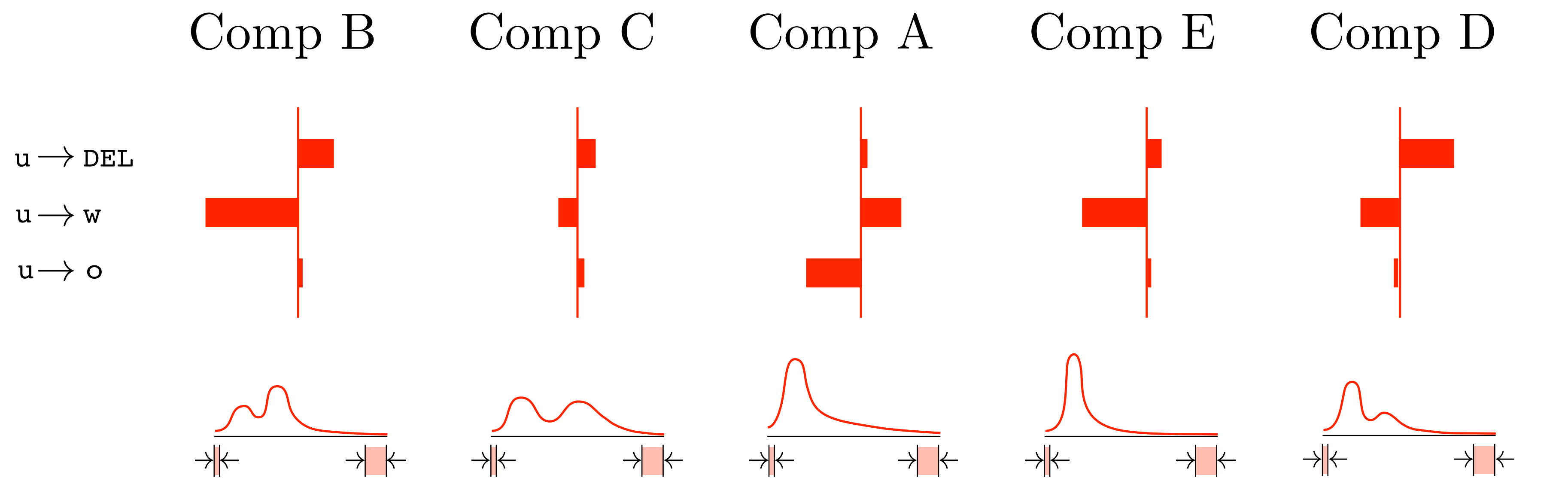
Feature model variant:
Includes individual edit operations as well as word spelling choices, incorporated using a log-linear parameterization. Whitespace lengths are modeled with separate multinomials for each compositor.

Experiments



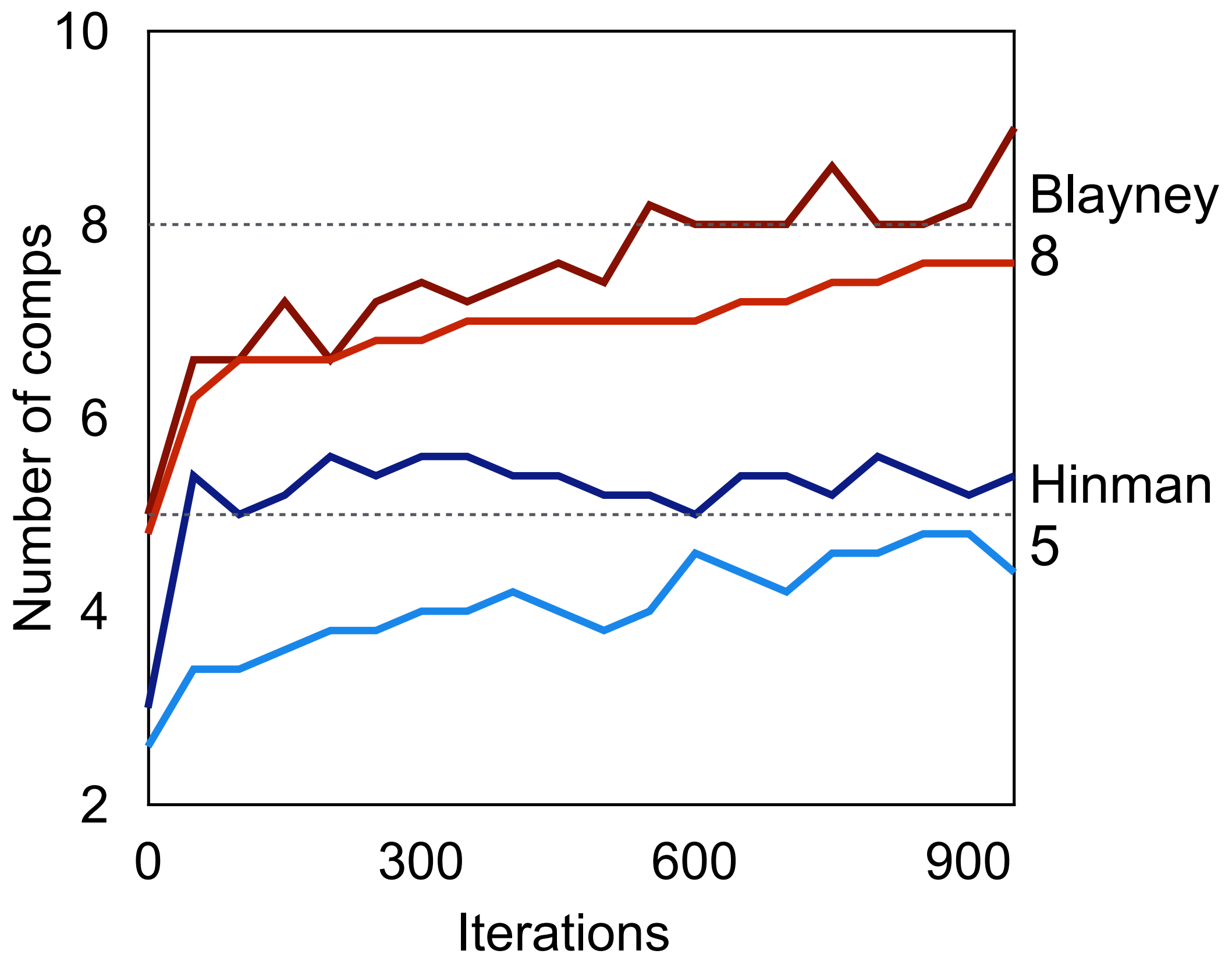
We evaluate by mapping the recovered page clusters to the gold compositors in the authoritative attribution and measure one-to-one accuracy. Including orthographic features substantially improves accuracy, but the best result is obtained by edit and whitespace features combined.

Analysis



Learned behaviors

Inspecting the parameters learned by our model reveals habits of individual compositors that have been noticed by bibliographers (Taylor, 1981).



Number of compositors

By extending our model with a non-parametric prior we are able to additionally learn the *number* of compositors. Depending on the subset of the vocabulary considered (e.g. words considered by Hinman vs. the larger set considered by Blayney) our non-parametric model agrees with the corresponding scholars' judgement.