Attributing the *Bixby Letter* using n-gram tracing

Jack Grieve (University of Birmingham, UK), Emily Carmody (Aston University, UK), Isobelle Clarke (Aston University, UK), Hannah Gideon (Aston University, UK), Annina Heini (Aston University, UK), Andrea Nini (University of Manchester, UK) and Emily Waibel (Aston University, UK)

In November 1864, only 5 months before he was assassinated, Abraham Lincoln, the 16th President of the United States, sent a short letter of condolence to the widow Lydia Bixby of Boston, the mother of five sons who were believed to have died in the Civil War. In fact, the Widow Bixby had only lost two sons and was also a Confederate sympathiser, who destroyed the letter in anger after reading it. Fortunately, the Adjutant General of Massachusetts, William Shouler, who had requested the letter from Lincoln in the first place, sent a copy to the *Boston Evening Transcript*, where the letter was published the following day. The *Bixby Letter* would go on to become one of America's most renowned pieces of correspondence and one of Lincoln's most celebrated texts, surpassed in notoriety only by the *Gettysburg Address* and the *Emancipation Proclamation*. It has also become part of public culture, for example being prominently featured in the movie *Saving Private Ryan* and being recited by President George W. Bush at the 10 year anniversary of the September 11th Attacks.

Despite its fame, the authorship of the letter has long been in dispute, with some historians claiming, based primarily on external evidence, that Lincoln's young personal assistant, John Hay, the future Secretary of State under McKinley and Roosevelt, was its author. Most notably, Burlingame (1995, 1999) argued that Hay was the author of the letter based on historical and linguistic evidence, including the use of the word *beguile,* which only Hay is known to have used in his writings. Overall, however, the debate is still unresolved and linguistic evidence has been minimal and inconclusive. The goal of this paper is therefore to investigate whether Lincoln or Hay is the more likely author of the *Bixby Letter* based on a quantitative comparison of the style of the letter to relatively large corpora consisting of the known writings of both statesmen.

First, all accessible writings of both Lincoln and Hay, which are voluminous especially for Lincoln, were obtained online from various historical archives. The texts were checked by hand to identify editorial notes and changes and to exclude co-authored and other problematic documents. We also choose to exclude all Lincoln's texts that were written after 18 May 1860, when Lincoln was nominated as the presidential candidate of the Republican Party, because Hay became Lincoln's personal secretary soon after that date. In total, the Lincoln contains corpus 1,085 texts totalling approximately 400,000 words, including substantial numbers of letters, speeches, bills, and resolutions. The Hay corpus contains 577 texts totalling approximately 260,000 words, including substantial numbers of letters, prose, poems, and diary entries.

This case of disputed authorship is especially challenging because the Bixby letter is so short. Totalling only 139 words, the Bixby letter is far too short to be attributed using standard stylometric techniques for authorship attribution, which generally requires disputed texts to be at least 500-1,000 words long so that the

relative frequencies of numerous linguistic features (e.g. function words) can be estimated accurately (Grieve, 2007). The usual approach to analysing short texts, for example in forensic linguistics where short texts are the norm, is therefore to look at whether or not forms in the questioned document occur in each possible author writing sample, ideally showing that the vast majority of linguistic forms found in the questioned text are only used by one possible author. There are, however, at least two major issues with this approach: how to select an unbiased feature set and how to control for variation in sample size.

Based on this general approach, and keeping these two limitations in mind, we have developed a new quantitative method for attributing short texts, which we refer to as *n-gram tracing*. The basic idea behind the method is to calculate the percentage of all n-grams in the questioned document that occur in each of the possible author writing samples. An n-gram is a sequence of one or more (e.g. 1-gram, 2- gram, 3-gram, etc.) linguistic forms in a text, which can be measured at any level (e.g. character-level, word-level). To conduct n-gram tracing, first all n-grams of a particular length and level (e.g. word-level 2-grams) are extracted from the questioned document. The percentage of those n-grams that occur in each of the possible author writing samples is then calculated. To control for variation in sample size, the percentages of forms are calculated for random samples of texts of different sizes drawn from each possible author writing. The possible author that uses a higher percentage of n-grams, especially as the size of these sample increases, is then selected as the most likely author of the disputed text.

To evaluate the method, we tested it using 1-4 word-level n-grams and 1-20 character-level n-grams on all 1,662 texts of known Lincoln and Hay authorship in our corpus. Specifically, we would remove one known text from our corpus, extract all the n-grams from that text, and then trace those n-grams across the remaining Hay and Lincoln texts to attribute the text. We have found the method to be highly accurate. Most notably, when attributing texts based on 4-10 character-level n-grams, the method attributes all 1,662 texts in the corpus of possible authors correctly. Remarkably, a majority of these texts contain fewer than 200 words and 10% contain fewer than 50.

Finally, to attribute the *Bixby Letter* we extracted all 1-4 word and 1-20 character level n-grams and then traced each set of n-grams across the Hay and Lincoln corpora. All analyses clearly identified Hay as the most likely author of the Bixby Letter. We therefore conclude that Hay as opposed to Lincoln was the author of the Bixby Letter, providing linguistic support for the analyses of Burlingame and others. N-gram tracing also appears to offer a solution to the problem of text length in authorship attribution, one of the most important problems in stylometry.

References

Burlingame, M. (1995). New Light on the Bixby Letter. *Journal of the Abraham Lincoln Association*, 16, 59-71.

Burlingame, M. (1999). The Trouble With the Bixby Letter: The stirring Civil War document featured in Saving Private Ryan grew out of a lie and probably wasn't really written by Lincoln. *American Heritage*, 50, 64-67.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22, 251-270.