Forensic linguistics: comparing word-length distributions using chi-square goodness-of-fit.

Author: Zachary Keyes * GitHub (https://github.com/imzoc/word-length-distribution-goodness-of-fit)

Introduction.

Forensic linguistics is involved with identifying authors based on their writing characteristics. Mendenhall claimed in 1887 that authors could be identified based on the word-length distribution in their writing [1]. However, while various comparisons in word-length distribution and mean were made, there were no supporting statistical probabilities or hypothesis tests in his paper. In this project, I will conduct a chi-square goodness-of-fit tests to test an initial null hypothesis:

H_0: Word-length distributions for all authors come from the same underlying distribution.
H_1: Word-length distribution for at least one author comes from a different underlying distribution than the other authors.

I will analyze three authors: Charles Dickens, Jane Austen, and Leo Tolstoy. I will use four books from each. For Charles Dickens, I will use Oliver Twist, Nicholas Nickelby, Great Expectations, and A Tale of Two Cities. For Jane Austen, I will use Pride and Prejudice, Emma, Persuasion, and Sense and Sensibility. For Leo Tolstoy, I will use War and Peace, Anna Karenina, Master and Man, and Resurrection. I cleaned all of the texts manually using automated Vim macros.

Hypothesis Tests and Results.

To test the first null hypothesis, we run hypothesis1() in main.py. This runs a chi-square goodness-of-fit test between the word-length distributions of all of the authors. The results can be seen in Table 1 and Figure 1.

Table 1: Chi-square goodness-of-fit results between authors.

Author	Chi-square	p-value
Jane Austen	345663.3553544331	0.0
Leo Tolstoy	372344.97461196635	0.0
Charles Dickens	685096.1621055813	0.0

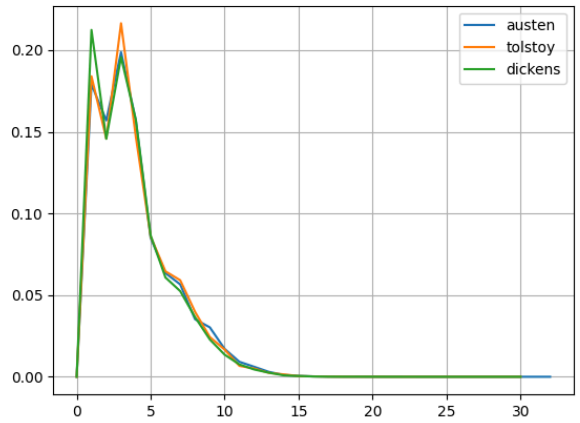

Figure 1: Comparing word-length distributions between authors.

As we can see, the likelihood that the three authors' underlying word-length distributions are the same is negligible. Each has a very high chi square value, especially Charles Dickens, who appears to have an even greater difference from the average across all three authors. We must therefore reject the null hypothesis and accept the alternative hypothesis that word-length distribution for at least one author (in this case, all of them) comes from a different distribution than the other authors.

What does this really mean? Just because each author has a different underlying word-length distribution doesn't mean that we can identify authors based on the word-length distribution of their work. If our goal is to identify a book's author based on the word-length distribution of their work,

we need to confirm that authors' works come from the same (or identifiably similar) word-length distribution. We need a second hypothesis:

H_{a0}: Word-length distributions for all books from an author come from the same underlying distribution.
H_{a1}: Word-length distribution for at least one book comes from a different underlying distribution than the rest of the books from that author.

To test the second null hypothesis, we run hypothesis2() in main.py. This runs chi-square goodness-of-fit tests between the word-length distributions of the books of each author (one author at a time. The results can be seen in Tables 2(a-c) and Figures 2(a-c).

The chi-square values in Tables 2(a-c), especially Table 2(a), are considerably less than those in Table 1, but it is clear that even books by the same author don't have the same underlying word-length distribution. We must reject the null hypothesis and accept the alternative hypothesis that the word-length distribution of at least one book (in this case all books) comes from a different distribution than the rest of the books from that author.

Table 2: Chi-square goodness-of-fit results between books by Jane Austen (a), Leo Tolstoy (b), and Charles Dickens (c).

Jane Austen (a)	Chi-square	p-value
Sense and Sensibility	50438.13111363382	0.0
Emma	37262.99123519585	0.0
Pride and Prejudice	49298.170266292705	0.0
Persuasion	70696.8037119288	0.0

Leo Tolstoy (b)	Chi-square	p-value
Resurrection	114815.20575019157	0.0
Anna Karenina	164637.68699546874	0.0
Master and Man	18380.225786761883	0.0
War and Peace	122164.35075348799	0.0

Charles Dickens (c)	Chi-square	p-value
Nicholas Nickelby	159478.89839345257	0.0
Oliver Twist	59124.26566772069	0.0
Great Expectations	92889.2417325738	0.0
A Tale of Two Cities	60098.019939543105	0.0

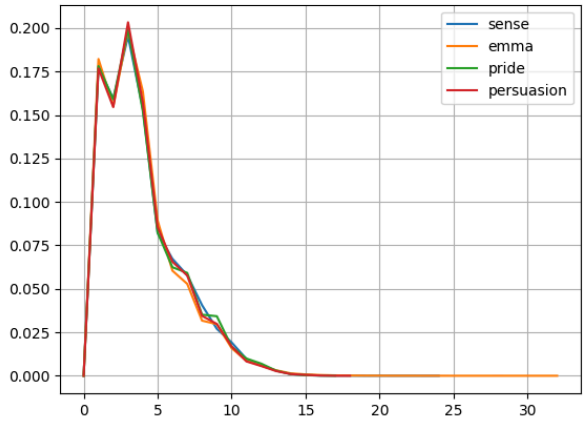

Figure 2a: Comparing word-length distributions between books by Jane Austen.

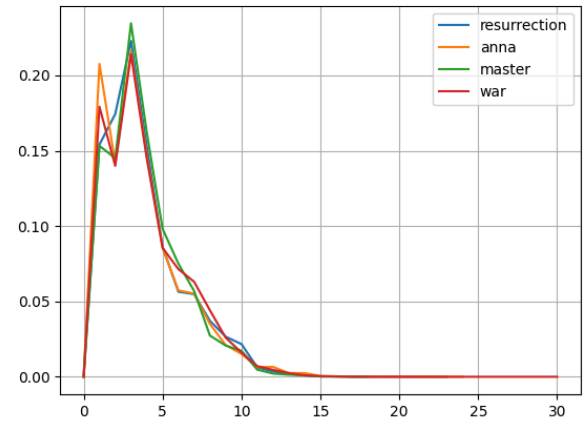

Figure 2b: Comparing word-length distributions between books by Leo Tolstoy.

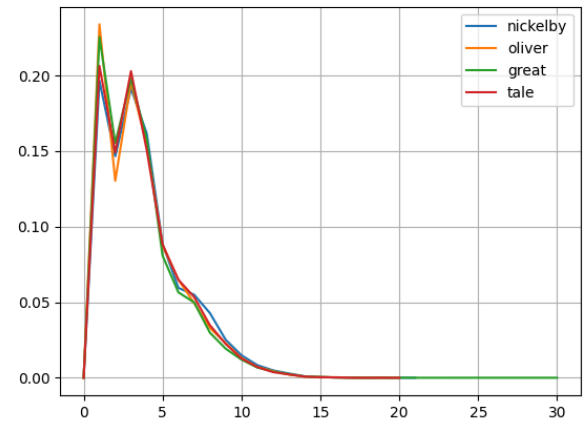

Figure 2c: Comparing word-length distributions between books by Charles Dickens.

Conclusion.

This article showed through chi-square goodness-of-fit tests that the word-length distributions of authors come from different underlying distributions. However, it also showed that the word-length distributions of books from the same author come from different underlying distributions. A book's word-length distribution is not enough to identify authorship using chi-square tests–not because authors have the same underlying word-length distribution, but because the underlying word-length distribution of works by the same author is different! This is not at all what I was expecting.

Future Directions.

It is important to note that the fact that word-length distributions of books by the same author come from different underlying distributions doesn't entail that the word-length distribution of a book can't be used to identify authorship. There may be more advanced statistical methods that could identify authorship based solely on word-length distribution, but that is beyond the scope of this project and a direction for future research.

References.

[1] Mendenhall, T. C. "The Characteristic Curves of Composition." *Science*, vol. 9, no. 214, 1887, pp. 237–49. *JSTOR*, http://www.jstor.org/stable/1764604. Accessed 6 May 2024.