# Authorship Discontinuities of El Ingenioso Hidalgo don Quijote de la Mancha as detected by Mixture-of-Experts

**Coufal, Christopher**
coufalc@duq.edu
Duquesne University

**Juola, Patrick**
juola@mathcs.duq.edu
Duquesne University

In the literary world, authorship of great novels is like writing a great piece of music; while there may never be a perfect way to determine if someone wrote a particular work or not, equations and algorithms have been developed in information theory and statistics to help those trying to discover the true authorship of contested written works. Because no method is perfect, using a set of methods on the same works can be used to give high probabilities of authorship. The JGAAP system houses a collection of methods such as Histogram Distance and Manhattan Distance and event sets such as word bigrams and character trigrams to allow users to perform multiple tests on contested works to see if the supposed author is actually the author by comparing samples from both the contested author and other possible authors.

We apply this framework and show a notable discontinuity in the authorial style of the novel *El Ingenioso Hidalgo don Quijote de la Mancha*, better known as *Don Quijote* (or *Don Quixote*).

## 1. Background

While there have been skeptics and scholars alike that have doubted Miguel de Cervantes Saavadre's true authorship of the entirety of *Don Quijote*, no one had tested whether or not Cervantes was in fact the true author of the whole of *Don Quijote*. The purpose of using the JGAAP system was to either give merit to or disprove this theory. By comparing to other authors who wrote works at about the same time *Don Quijote* was written, the JGAAP system would test to see if the text that Cervantes supposedly wrote was closer to the first volume of *Don Quijote* or closer to other authors of the same time period. If a definitive break could be established between where the program attributed Cervantes as the author and where it did not, that would suggest either a major style shift or the presence of another author different from Cervantes, while no break at all would suggest that Cervantes was in fact the true author of the second volume of *Don Quijote*, assuming that he was also the author of the first volume.

## 2. Methods and Materials

For this authorship attribution, the program JGAAP 4.0 was downloaded from `http://www.jgaap.com`, developed by Patrick Juola at Duquesne University. The *Don Quijote* text used was acquired from Project Gutenburg at `http://www.projectgutenburg.org`. The full text of Don Quijote was then stripped of the introductions and separated into chapters by volume. The first volume was then set as the basis for Miguel de Cervantes' original authorship. Every third chapter, starting with chapter three, was used as the base case for Cervantes' work. Two other authors used for comparison, Fransisco de Quevedo and Mateo Alemán, were also used. Quevedo's, *Historia de la vida del Buscón, llamado Don Pablos, ejemplo de vagamundos y espejo de tacaños and Alemán's Guzmán del Alfarache* were also taken from Project Gutenburg and broken into roughly the same number of chapter-type sections as the number of chapters used for Cervantes' *Don Quijote*. In order to make sure that Cervantes' was actually the author of volume one of *Don Quijote*, every chapter not used in the base case was compared to the base chapters, Quevedo's work, and Alemán's work. Each test used JGAAP's Normalize Whitespace, Strip Punctuation, and Unify Case canonicizers on all of the documents. Five event sets - Word, WordBiGram, WordTriGram, WordTetraGram, and Word Length - were all paired with nine analysis methods - Camberra Distance, Cosine Distance, RN Cross Entropy, Histogram Distance, Kullback Leibler Divergence, Levenshtein Distance, Manhattan Distance, KS Distance, and Naive Bayes

Classifier, for a total of 45 unique event set-analysis methods. Once the first volume of Cervantes' work was confirmed to be uniformly Cervantes', volume two of *Don Quijote* was tested in the same manner as the first volume in order to provide an accurate analysis.

We apply a mixture-of-experts approach to the evaluation of authorship. Each different method is treated as a single "expert" in different aspects of authorial style, and permitted to vote on who (among the candidates) is the author of any specific fragment. If all 45 test "experts" vote on Cervantes, we consider this to be strong evidence supporting his authorship, while if only 5 or so of the 45 consider Cervantes to be the most likely author, we consider this to be evidence *against*.

## 3. Results

As a result of the analysis on the second volume of *Don Quijote*, the JGAAP program indicated that starting at chapter 6 Cervantes was not the author. Out of the 45 tests run on each chapter, the chapters in the first volume had a mean of 37.54 occurrences of Cervantes as the author with a standard deviation of .852. The first five chapters of the second volume had a mean of 36.50 occurrences of Cervantes as the author with a standard deviation of 1.517. Chapters 6-74 of the second volume, however, had a mean of 4.90 occurrences of Cervantes as the author with a standard deviation of 1.436. This radical shift in authorship means either Cervantes completely shifted his writing technique or he did not write the latter 69 chapters of the second volume of *Don Quijote*.

## 4. Discussion

While there are people who are skeptic about the authorship of *Don Quijote*, nothing up until now has given those claims any grounds other than speculations based on inconsistencies in the text. Although this analysis does not guarantee that Cervantes did not write the last 69 chapters of the second volume, it does make the probability of that claim much greater. This, in part, is due to the fact that none of the tests in JGAAP has been tested enough to show that it will work for all documents. As further analysis of the methods continues, the results of the tests used in this authorship attribution will most likely validate these results. As tests and methods prove to not work, the analysis will be redone with these tests omitted from the analysis giving a more accurate result.