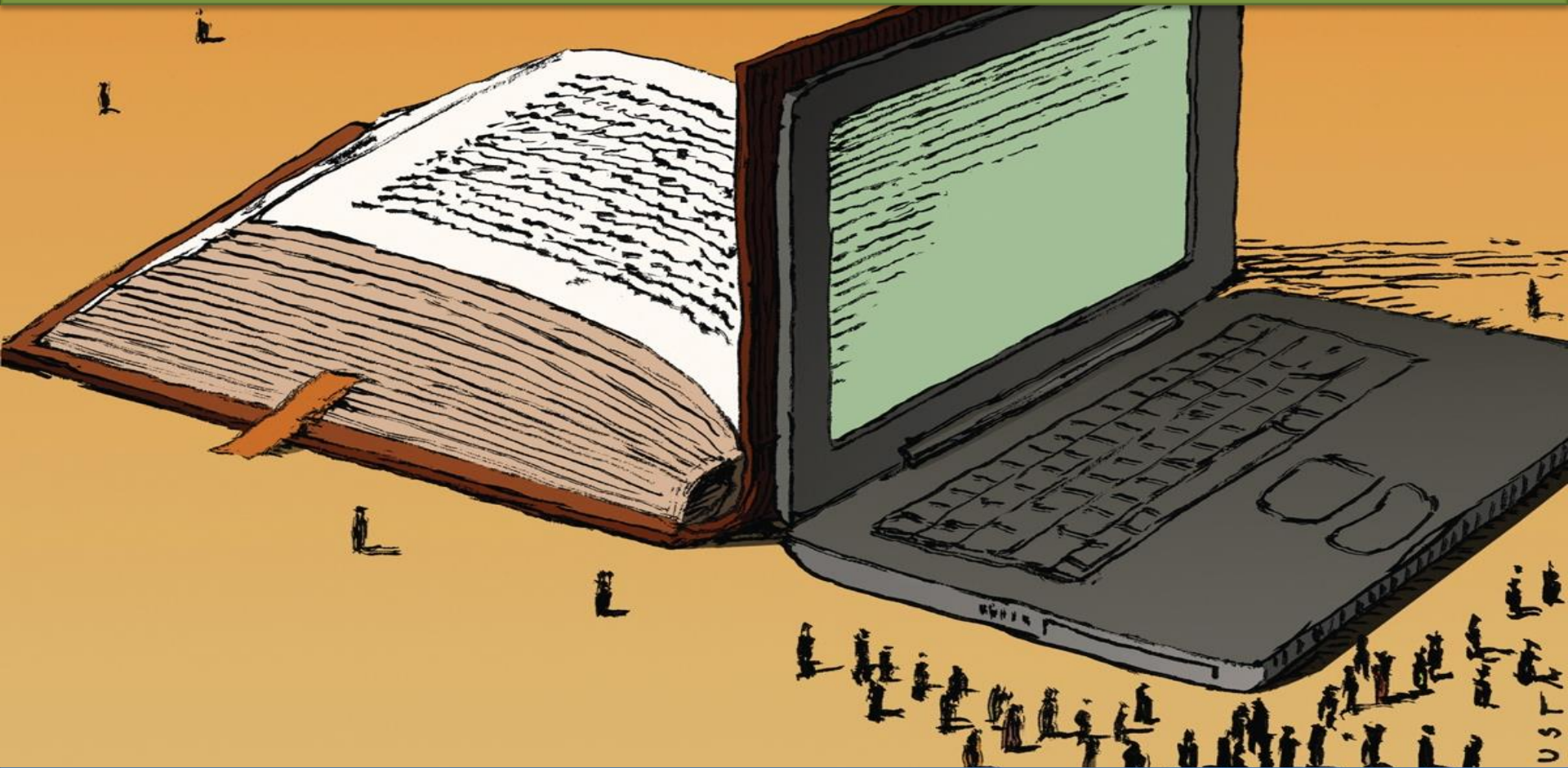


Stylometry



Literature in the Digital Age (Digital Humanities)

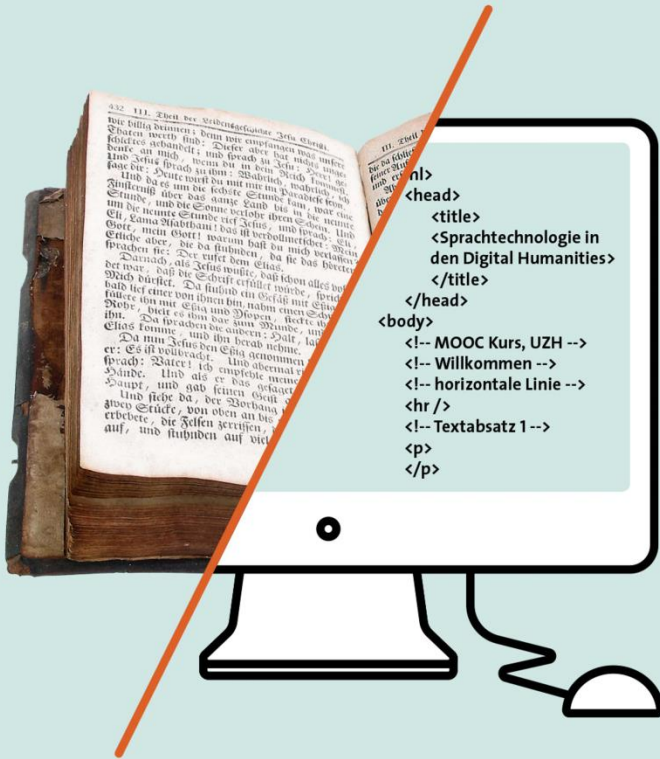
Overview

- Background to Stylometry
- What are Digital Humanities ?
- Behavioural Biometrics
- What is Stylometry ?
- Origin & History of Stylometry
- How did Stylometry gained popularity ?
- Applications of Stylometry
- Close Reading & Distant Reading
- Principles of Stylometric Analysis
- Stylometric Analysis
- Stylometric Software Packages

What are digital Humanities ?

Digital Humanities

- It refers to an academic field concerned with the application of computational tools and methods to traditional humanities disciplines such as literature, history, and philosophy.



Origin & History of Stylometry

1. Stylometry dates back to 15th century and the comparison of differing translations.
2. In 1851, August de Morgan, a British mathematician, speculated that word length might prove to be a unique marker of an author's style.
3. The beginning of stylometry is generally traced to the early suggestion of resolving authorship disputes by Augustus de Morgan through the frequency of word lengths in 1851.

What do the names, Sir Walter Raleigh, the Earl of Oxford, Christopher Marlowe and Queen Elizabeth have in common?

Each has been cited as the possible author of the plays attributed to William Shakespeare. It's an historical who-done-it.

How did Stylometry Gained Popularity ?

- Stylometry gained popularity from renaissance drama authorship questions. This interest in determining authorship in renaissance works helped to establish stylometry's credibility in the area of authorship attributions.
- Augustus de Morgan's suggestions prompted the first manual quantitative analysis in the late 1880s by Thomas C. Mendenhall who used word length distributions from the works of Bacon, Marlowe, and Shakespeare for identifying the true author of Shakespeare plays.

Origin & History of Stylometry

The basics of stylometry were established by the Polish philosopher Wincenty Lutoslawski in 1890. He published *Principes de stylométrie* in 1890, thus coining the term “stylometry.”

Many years passed before George Kingsley Zipf discovered a relationship between the rank and frequency of words in 1932, which later came to be known as Zipf's Law. Similar efforts followed, such as George Yule's measurement of word frequency for analysis of vocabulary richness in 1944, which is now known as Yule's Characteristic. However, the research literature largely refers to the work of Mosteller and Wallace on the Federalist Papers in the early 1960s as the foundation of computer-assisted stylometry [117–119], while the Federalist Papers remain as a corpus of interest .

What is Stylometry ?

- Stylometry is the quantitative study of literary style through computational distant reading methods. It is based on the observation that authors tend to write in relatively consistent, recognizable and unique ways.
- The analysis of authorial style, termed stylometry, assumes that style is quantifiably measurable for evaluation of distinctive qualities.

Examples of unique features of Styles

- Each person has their own unique vocabulary, sometimes rich, sometimes limited. Although a larger vocabulary is usually associated with literary quality, this is not always the case. Ernest Hemingway is famous for using a surprisingly small number of different words in his writing,¹ which did not prevent him from winning the Nobel Prize for Literature in 1954.
- Some people write in short sentences, while others prefer long blocks of text consisting of many clauses.
- No two people use semicolons, em-dashes, and other forms of punctuation in the exact same way.

The Core Aspect of Stylometry

- As David Holmes explains , “At [stylometry’s] heart lies an assumption that authors have an unconscious aspect to their style, an aspect which cannot consciously be manipulated but which possesses features which are quantifiable and which may be distinctive.”

Strategies in Stylometry

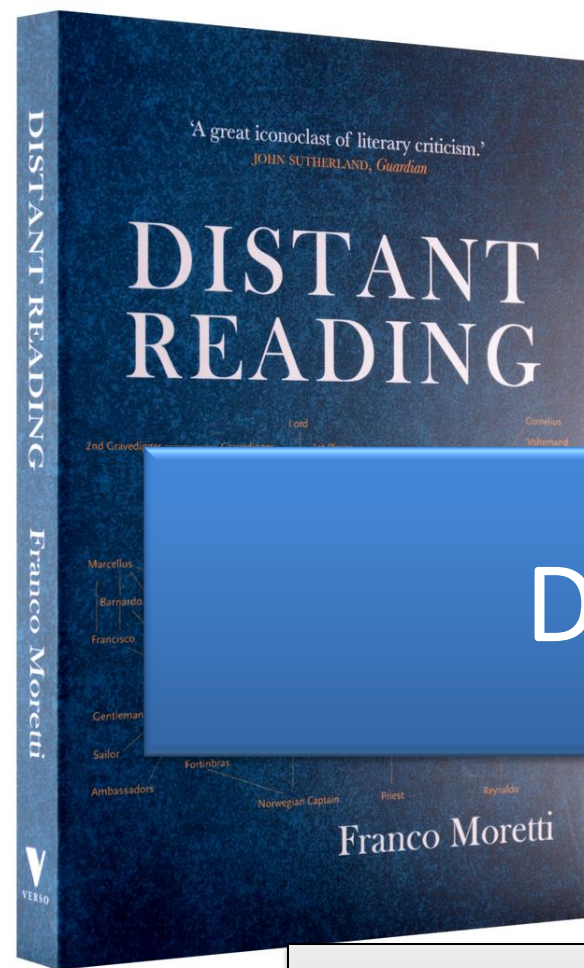
- Like many strategies in the digital humanities, stylometry combines traditional *close reading* alongside more *distant reading*.

Close Reading

In literary criticism, close reading describes a sustained attention to the text, looking at word choice, syntax, and particular images, to reveal meaning. This methodology came out of New Criticism and does not rely on historical or biographical research, intending to analyze the complexities of the individual text instead.

beatle, she didn't
myself, and thro
ultimate trip.⁸⁰

Freedom is in the m
the man's suppose
supposed to be som
relationship. It ju
past they work
digging the
they'd spl
should



DISTANT READING

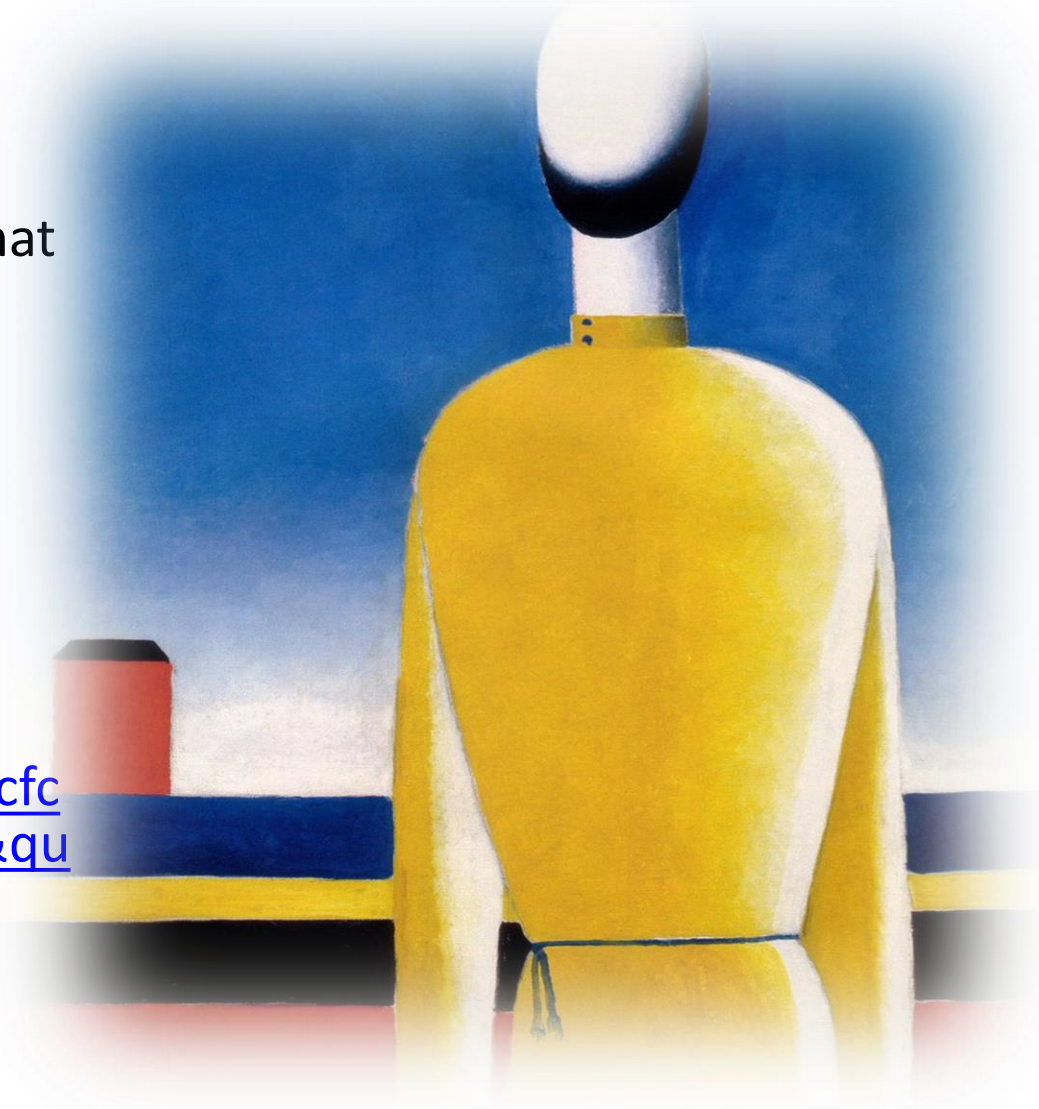
‘Distant Reading’ is a phrase coined by the literary critic Franco Moretti (2003)

Distant reading

Distant reading refers to a professional reading method that relies heavily on computer programs. It uses big data analytics for the purposes of literary scholarship. This is a quantitative technique .

Here is an example

- <https://voyant-tools.org/?corpus=4297253cfc2bd73976545d28864cc2f4&query=like>



Distant reading

- This is a concept coined by Franco Moretti that suggests that looking at the wider scope of literature, through larger computational or archival methods, can help us see larger trends and reveal previously obscured systems within literary study.

Close Reading & Distant Reading

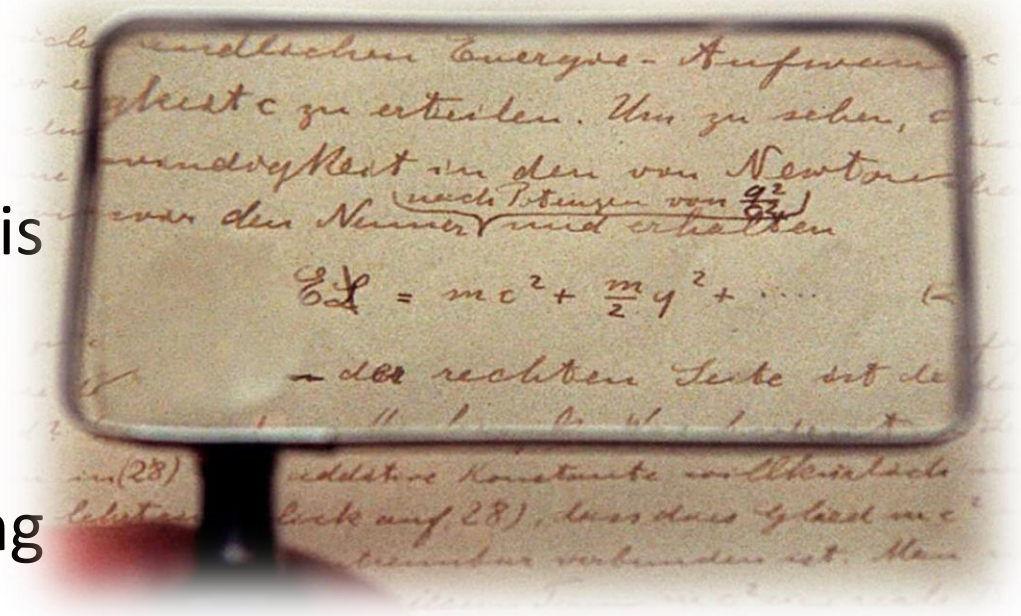
- In close reading a reader embarks on painstaking analyses of the semantic and syntactic intricacies of single literary texts.
- Narrow range of literary text selected
- In distant reading, researcher mines huge databases that contain thousands of literary texts, to identify recurring patterns and large scale historical developments across national borders, and over whole centuries. One such database that Moretti and his fellow researchers used, and that you can use, is Google's Ngram database.
- Thousands of literary texts selected

Benefits of close reading

- Identify Large patterns prevailing in an age
- Open up the canon
- Greater Objectivity is ensured
- (Traditional literary scholarship tends to be subjective in the end, shaped by the literary scholar's own norms, values, and prejudices).
- By using the methods from the social sciences and the natural sciences, and by embarking on quantitative analysis of big data, distant reading promises to give us greater objectivity and greater comprehensiveness.

Stylometry Applications

1. Plagiarism Checker
2. Forensic analysis
3. Author Identification
4. Software Code analysis
5. Genre classification
6. Historical Study
7. Context Understanding
8. Music Lyrics
9. Painting



Applications of Stylometry

- Scholars have used stylometry as a tool to study a variety of cultural questions. For example, a considerable amount of research has studied the differences between the ways in which men and women write or are written about.



-

Applications of Stylometry 2

- However, one of the most common applications of stylometry is in authorship attribution. Given an anonymous text, it is sometimes possible to guess who wrote it by measuring certain features, like the average number of words per sentence or the propensity of the author to use “while” instead of “whilst”, and comparing the measurements with other texts written by the suspected author.

What is Authorship Recognition?

The basic question: “who wrote this document?”

Stylometry can identify authors based on their writing.



Pakistan Peoples Party

To the officials and members of Pakistan Peoples Party I say that I was honoured to lead you. No leader could be as proud of their party, their dedication, devotion and discipline to the mission of Quaid-e-Azam Zulfikar Ali Bhutto for a Federal, Democratic and Egalitarian Pakistan as I have been proud of you. I salute your courage and your sense of honour. I salute you for standing by your sister through two military dictatorships.

I fear for the future of Pakistan. Please continue the fight against extremism, dictatorship, poverty and ignorance.

I would like my husband Asif Ali Zardari to lead you in this interim period until you and he decide what is best. I say this because he is a man of courage and honour. He spent 11 1/2 years in prison without bending despite torture. He has the political stature to keep our party united.

I wish all of you success in fulfilling the manifesto of your party and in serving the down-trodden, discriminated and oppressed people of Pakistan. Dedicate yourselves to freeing them from poverty and backwardness as you have done in the past.

Benzvi Bhutto

October 16, 2007

Application of Stylometry in Crimes

- From the Institute for Linguistic Evidence:
- “In some criminal, civil, and security matters, language can be evidence... When you are faced with a suspicious document, whether you need to know who wrote it, or if it is a real threat or real suicide note, or if it is too close for comfort to some other document, you need reliable, validated methods.”

Application of Stylometry in Identification of Authorship

- Scenario: Anonymous Forum vs. Oppressive Government.
- Participants organize protests.
- Posts are completely unlabeled (no pseudonyms)
- Unknown organizational structure, number of authors, etc.
- The government applies unsupervised stylometric techniques.
- # of authors may be discovered, author profiles created.
- Fed into supervised stylometry system to identify individuals.

- The key concept behind author identification is the process of feature engineering, where the machine selects features from the collected text that suitably describes the style of an individual author which helps in distinguishing the author from other writers.

Writing Style analysis Assumptions

- “There is an unconscious aspect to an author’s style that cannot be consciously manipulated but which possesses quantifiable and distinctive features”
- Researchers believe typical individual human activities carry invariant similarities and slightly vary from one person to another. Similarly, the style in writing is usually distinguished by the repeated choices of words or text patterns that the writer tends to make subconsciously. These repeated choices are very individualistic and are supposed to reflect a writer’s style.

Principles of Stylometric Analysis

- Stylometry looks at a variety of features of an author's style such as:
- Word length
- Sentence length
- Paragraph length
- Punctuation
- Function words (for example: the, and, a, of, to, in, that, with, but, it)
- Letters
- N-grams, bigrams, trigrams (characters in a row)
- Bi-words and Tri-words (two or three words occurring in a certain order)

Stylometry Analysis in the Past & Present

- In the past, this discipline was very limited because humans had to manually perform all of the data analysis on the texts they were studying. With the advent of computers, however, researchers are freed from data analysis roles to focus on the theory underlying the data relationships. Stylometry research has yielded several methods and tools over the past 200 years to handle a variety of challenging cases.

-

Stylometric Analysis

- Before you are able to input any of your texts into various stylometric programs, it's important that you take a step back and prepare your corpus by collecting all of your texts, organizing them into readable files, and removing any metadata.
- As the analysis gets more and more thorough and complex, professionals move away from this judgement process and end up with tools that allow them to make confident claims about Authorship and style . These tools take the form of statistical tests that tell us about a document as well as software to facilitate the combination of multiple statistical tests (for example , JGAAP, and Stylo)

Software Packages for Stylometry

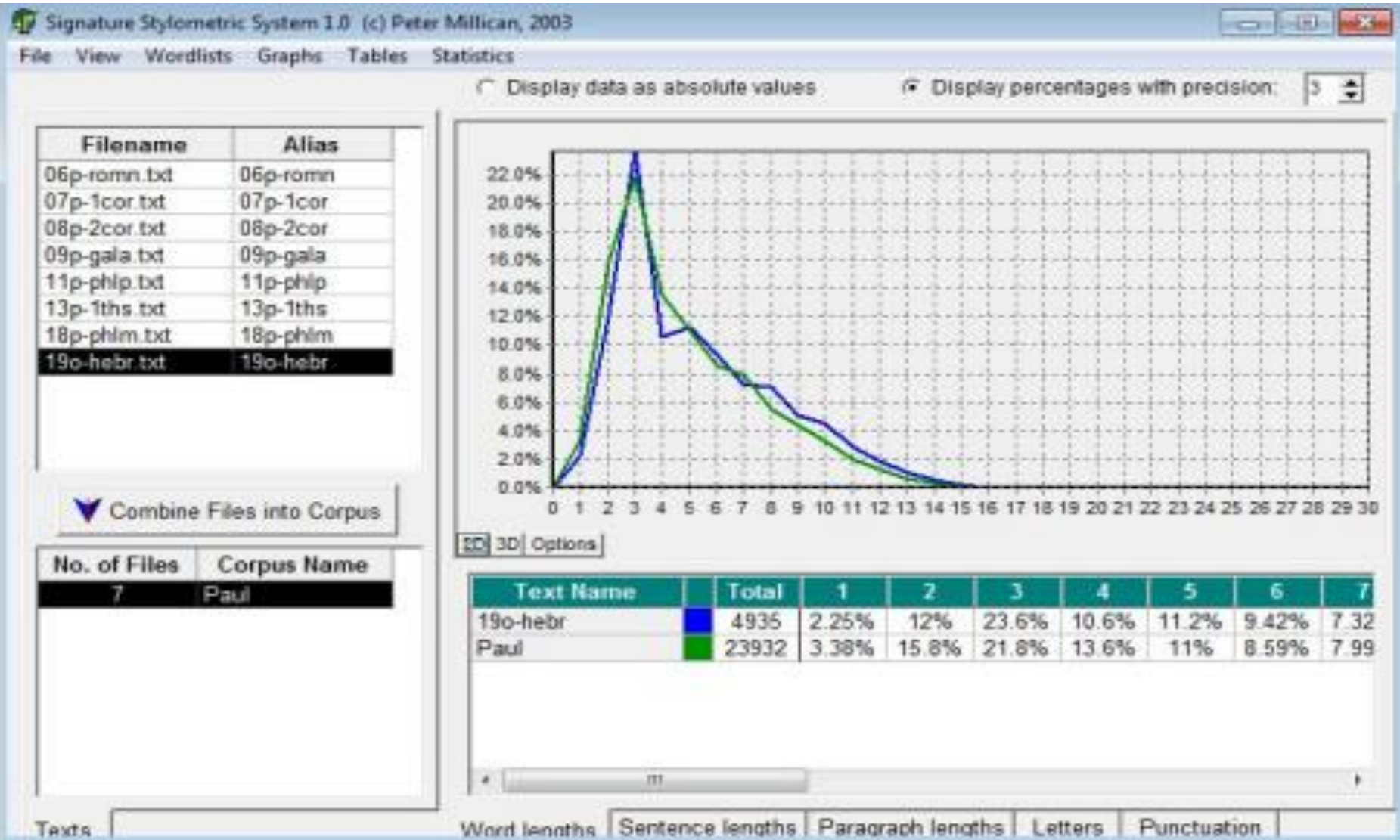
Three prominent tool packages for do-it-yourself stylometry:

- Signature Stylometric System
- Java Graphical Authorship Attribution Program (JGAAP)
- The R library Stylo.

Signature (Software) Stylometric System

- The software Signature is the most basic, and offers simple ways to analyze documents by directly comparing properties against each other. Signature allows you to compare word length, sentence length, paragraph length and punctuation and letter frequency. Based on these attributes, Signature generates a graph comparing numerous documents with the option of measuring values as percents for standardization or doing a chisquare test on two documents.
- This is less useful than it sounds though—there is no documentation included explaining how keywords are ordered and signature can only graph userselected words leaving the user with no guidelines on how to create meaningful word comparisons

Signature Stylometric System showing sentence length of two books



Limitations of Stylometry

- An author has free will to purposely alter one's style according to genre, topic, manuscript guidelines, and so on. The ability to effectively write in several contexts is likely beneficial for a successful writer, but this same inconsistency further complicates authorship analysis.

Limitations of Stylometry

- Moreover, stylometry does not presently offer studies on the general population for making comprehensive conclusions regarding authorial style. As a result, most of the information gained from a single study is isolated to the researcher's dataset. Furthermore, as author sets become larger, stylometry features lose discriminating power; thus, stylometry characteristics do not carry the same identifying capabilities as physiological biometrics, as is generally the case with behavioral modalities.

Conclusion

- **Stylometry** is the use of statistical analysis of style. In other words, the quantification and computation of style, is called **stylometry**. **stylometry** relies on digitized versions of texts. **Stylometry** tests for author authenticity, plagiarism, multiple authors, and other factors. It is commonly employed in universities to test for student's theses plagiarism. It has also been used in high-profile matters such as authorship identification.

Thank you very much