# Stylometry, part 2

## Unmaksing J.K. Rowling

novels by 4 other, contemporary female authors from the UK, writing serial fantasy: Diana Wynne Jones, Jenny Nimmo, Susan Cooper, Edith Nesbit

This makes them a good control group for our analysis. You don't want to compare Rowling to authors who write in a completely different genre or style.

## Dendrogram

A dendrogram is a diagram that shows the arrangement of the clusters produced by hierarchical clustering. It is a tree-like structure that represents the relationships between the clusters and the data points. The dendrogram is used to visualize the results of the clustering process and to help identify the optimal number of clusters.

You read it from left to right, and the height of the branches represents the distance between the clusters. The longer the branch, the more dissimilar the clusters are.

## Include Galbraith

Other options: Via **culling**, users can specify the percentage of samples in which a feature should be present in the corpus in order to be included in the analysis. Words that do not occur in at least the specified proportion of the samples in the corpus will be ignored.

## Bag of words and z-scores

The starting point for the document representation is a **'bag of words' model** of the text, i.e. we count how often each word form occurs in each document. The word counts are then transformed to relative frequencies to compensate for different text lengths. For further processing, the n most frequent different words over the whole corpus (hereafter, nMFW) are chosen.

In the vector space model, each different word corresponds to a different dimension. The word frequencies of all documents can now be arranged in a documents words matrix.

Following Burrows' approach: the word frequencies are standardized, i.e. they are normalized so that, over the whole corpus, **the mean for each word is 0 and the standard deviation is 1** (the result is also known as the 'z-score').

**This reduces the influence of the top-scoring words**: Since word frequencies follow the distribution described by Zipf's law (published in 1935), the distance would otherwise barely be influenced by anything but a few top-scoring words.

A **z-score is a statistical measurement that describes a value's relationship to the mean of a group of values**. It is used in stylometry as a standardized way to compare the frequency of a specific feature, like a word, in different texts or corpora.

The z-score is measured in terms of standard deviations from the mean. If a z-score is 0, it indicates that the data point's score is identical to the mean score. A z-score of 1.0 would denote a value that is one standard deviation from the mean. Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

Here's the mathematical formula for calculating a z-score:

```
Z = (X − μ) / σ
```

Where:

- Z is the z-score
- X is the value of the element
- μ (mu) is the population mean
- σ (sigma) is the standard deviation

## Distance metrics

The **'Manhattan distance'** (L1 norm of the difference vector) sums up the absolute distances between each word's normalized frequencies in the two documents. Manhattan Distance sums the absolute differences between the coordinates of the vectors. It might be more suitable for certain textual features that do not conform to Euclidean geometry.

The **'Euclidean distance'** (the L2 norm of the difference vector) calculates the 'straight line' distance between the vectors. Euclidean Distance measures the straight-line distance between two points (or vectors) in the vector space. In stylometry, it can quantify how far apart two authors' stylistic word usage patterns are.

The **'cosine distance'** corresponds to the angle u between the vectors. Cosine Similarity measures the cosine of the angle between two vectors, effectively capturing the orientation (pattern similarity) rather than the magnitude difference. It's particularly useful in stylometry for comparing the overall stylistic direction of authors, regardless of their verbosity.

Reference: Evert, Stefan, et al. "Understanding and Explaining Delta Measures for Authorship Attribution." Digital Scholarship in the Humanities, vol. 32, no. Supplement 2, Dec. 2017, pp. ii4–16. https://doi.org/10.1093/llc/fqx023.

e.g. text 1 can be represented as a vector with these coordinates: (0.5, 0.3, 0.2, 0.1); text 2 as (0.6, 0.2, 0.1, 0.1). These vectors represent the frequencies of ONLY four words in the texts! The distance metric (e.g., Euclidean) will calculate the distance between these two vectors. The result will be a single number that represents the distance between the two texts. The smaller the number, the more similar the texts are in terms of the frequencies of these four words.

## Principal Component Analysis (PCA)

Next, these distances are subjected to **Principal Component Analysis (PCA)**. This means that the distances are transformed into a new set of distances that are linear combinations of the original distances.

PCA is essentially a dimensionality reduction technique that transforms a large set of variables (in this case, words or features from the BoW model) into a smaller one that still contains most of the original data's variation. In stylometry, PCA helps by reducing the high-dimensional BoW vectors into a lower-dimensional space, where stylistic patterns become more discernible.

This reduced space can reveal clusters of texts or authors, indicating similarities in their stylistic features. By plotting the principal components, researchers can visually inspect and interpret the relationships between different authors or texts, often uncovering underlying stylistic dimensions (e.g., formality, complexity).

PCA will then transform this vector into a new one with fewer dimensions, e.g. (0.7, 0.1). The new vector retains most of the original information, but in a more compact form.

## Caesar

The 'War Commentaries' by Julius Caesar (Corpus Caesarianum) refers to a group of Latin prose commentaries, describing the military campaigns of the world-renowned statesman Julius Caesar (100–44 BC), the founder of the Roman Empire. While Caesar must have authored a significant portion of these commentaries himself, the exact delineation of his contribution to this important corpus remains a controversial matter. Most notably, Aulus Hirtius –oneof Caesar's most trusted generals –issometimes believed to have contributed significantly to the corpus.

The Caesarian Corpus is composed of five commentaries describing Caesar's military campaigns:

- Gallic War Bellum Gallicum, conquest of Gaul, 58–50 BC;
- Civil War Bellum civile, civil war with Pompey, 49–48 BC;
- Alexandrian War Bellum Alexandrinum, Middle East campaigns, 48–47 BC;
- African War Bellum Africum, war in North Africa, 47 to 46 BC
- Spanish War Bellum Hispaniense, rebellion in Spain, 46–45 BC.

The first two commentaries are mainly by Caesar himself, the only exception being the final part of the Gallic War (Book 8), which is commonly attributed to Caesar's general Aulus Hirtius (c90 – 43 BC). **Caesar's primary authorship of these two works, except for Book 8, is guaranteed** by the ancient testimonia of **Cicero**, **Hirtius**, **Suetonius**, and **Priscian** as well as the unanimous evidence of the manuscript tradition.

**Caesar's ancient biographer Suetonius**, writing a century and a half after his death, suggests that **either Hirtius or another general, named Oppius**, authored the remaining works: *'[Caesar] also left commentarii of his deeds during the Gallic War and the Civil War with Pompey. For the author of the Bellum Alexandrinum, Africum, and Hispaniense is uncertain. Some think it is Oppius, others Hirtius, who supplemented the last, incomplete book of the Bellum Gallicum'*

We also have **a letter of Hirtius to Cornelius Balbus**, a fellow supporter of Caesar, which is transmitted in the manuscripts preceding the Hirtian 8th book of the Gallic War. In this letter, Hirtius lays out his project:

*'I have continued the accounts of out Caesar on his deeds in Gall, since his earlier and later writings did not fit together, and I have also finished the most recent and incomplete account, extending it from the deeds in Alexandria down to the end, not admittedly of civil discord, of which we seen no end, but of Caesar's life'*

Despite occasional doubts, the most recent analysis has shown that there is no reason at all for doubt ing the authenticity of the letter.

Hence, a puzzle that has persisted for nineteen centuries: what are the relationships of the different war commentaries to one another, to Hirtius, and to Caesar?

Our analyses broadly support the following conclusions:

1. **Caesar himself wrote, in addition to Gallic Wars, books 1–7 and the Civil War**, as well as the first 21 chapters of the Alexandrian War.
2. **Hirtius wrote Book 8 of the Gallic Wars** and the remainder of the Alexandrian War.
3. At least **one other author wrote the African War and the Spanish War**. The African War and the Spanish War were probably written by two different authors.
4. Our results do not invalidate Hirtius's own claim that he himself compiled and edited the corpus of the non-Caesarian commentaries.
5. The significant **stylistic heterogeneity** we have detected in parts of the Gallic War and the Civil War likely represents Caesar's compositional practice of relying on, and sometimes incorporating, the briefs written for him by his legates.

## Hildegard of Bingen

The **Benedictine nun Hildegard of Bingen** was one of the most productive female authors of the Middle Ages. After a youth as anchoress at the abbey of the monks of Disibodenberg in the Rhineland near Mainz, she ended up as **abbess** of **her own convent at the nearby Rupertsberg**.

Her extensive oeuvre includes genres as diverse as **visionary books, letters, hagiographical texts, treatises on monastic life, musical compositions, and some works on physics and medical healing.**

Considered a true **prophetess**, receiving revelations and admonitions from God, she enjoyed a special status, even in the highest ecclesiastical milieux.

Her **extensive circle of correspondents**, comprising, among others, **popes and the emperor**, testifies to her prophetic reputation. She was therefore able to gain an **authority unprecedented for a woman**, enabling her to even criticize the male clergy of her time.

Our modern post-romantic conception of authorship therefore seems profoundly anachronistic with respect to the Middle Ages. Yet, even if medieval culture did not share our present-day view on the significance of original authorship, the Middle Ages have known many respected and authoritative individuals who were recognized by their contemporaries and posterior readers as producers of very specific literary works. Some kind of correlation even existed between the degree to which texts were susceptible to alterations and the religious and intellectual authority of their authors.

This did not mean, however, that such recognized authors were necessarily acting individually in the process of conceiving their treatises or narratives—quite the contrary. **Writing in the Middle Ages meant entering into a dialogue with a long line of predecessors, whether through citations, paraphrasing, or allusions.** In the actual process of literary composition too, medieval authors only seldom worked alone.

Women writers like the German nuns Hildegard of Bingen (1098–1179) or Elizabeth of Schonau (1129–1165) **were considered unlearned and incapable of independently writing down their visionary experiences**,

even if these were 'divinely inspired'. These women therefore had to be assisted by male collaborators, often also serving as their spiritual directors.

The precise nature and implications of such cross-gender collaborations remain a topic of scholarly debate.

In one of her vitae, **her biographer Guibert of Gembloux** specifies that she was 'uneducated as to her schooling in the art of grammar' (Derolez, 1988–1989, p. 377).

At the very end of her life, however, she was unexpectedly joined by Guibert, a monk from the abbey of Gembloux in Brabant (nowadays Belgium). Himself a fervent letter writer and hagiographer, he served as her **secretary from 1177 until her death in 1179**.

The question has been raised **to what extent Hildegard's secretaries interfered with the final versions of her works, possibly generating male, clerical interpretations rather than original female viewpoints.**

The immediate incentive for the present article is the preparation of a new critical edition of two lesser known texts attributed to Hildegard of Bingen, supposedly dating from the last years of her life:

- the *Visio de Sancto Martino*, which is conceived as a letter addressed to the worshippers of Saint Martin,
- and the *Visio ad Guibertum missa*, containing spiritual advice to an anonymous monk-priest, generally identified as her last secretary, Guibert of Gembloux (1124–1213).

Among the few scholars who paid attention to these texts, **there is still no consensus as to the extent to which they should be attributed to either Hildegard herself or to her collaborator Guibert**. As neither traditional stylistic analysis nor contextual historical research has so far been able to resolve the problem, we will approach this issue through a stylometric analysis.

First, **does stylometry allow for an authorial differentiation** between the writings of twelfth-century Latin authors, belonging to highly similar intellectual circles? To answer this question, we will investigate the letter collections or epistolaria of Hildegard of Bingen, her secretary Guibert of Gembloux, and their **famous contemporary, Bernard of Clairvaux**.

Our aim is to assess **to what extent we can distinguish stylistic profiles for these authors**, despite the marked variance within medieval manuscript culture (Cerquiglini, 1999), as well as **the fact that these authors, like many of their contemporaries, were often assisted by secretaries.**

Next, we wish to analyze in more detail to what extent we can discern in Hildegard's epistolary work, the influence of her last secretary, Guibert of Gembloux. **Did her style undergo detectable stylistic changes under the editorial assistance of Guibert, or does the same homogeneous authorial voice appear throughout her epistolary work?**

The **Visio de sancto Martino** ('Vision of Saint Martin') and **Visio ad Guibertum missa** ('Vision sent to Guibert'), which are at stake in this article, cannot be found in the Riesenkodex. They are only preserved in three manuscripts that can be linked to the abbey of Gembloux and Guibert's own oeuvre. Therefore, both texts are traditionally not included in the core of Hildegard's canon.

Whereas the titles in the manuscripts (Fig. 2), as well as Guibert's accompanying letters, firmly attribute these visiones to Hildegard, **there are good reasons to suspect that Guibert must have been extensively involved in their final redaction.**

- The figure of Saint Martin for instance—the main topic of the *Visio de sancto Martino*—is entirely absent from Hildegard's oeuvre. Guibert, on the other hand, developed a lifelong fascination for this saint and devoted nearly half of his life to spreading his cult. The Visio ad Guibertum missa discusses the role of the priest as well as the topic of literary collaboration, both issues of direct relevance to Guibert.

- Moreover, the end of the latter text contains a passage of particular interest in which Hildegard grants Guibert the exceptional right to revise her texts more fundamentally than simply at the level of style and grammar:

  *With this statement, Hildegard allegedly granted Guibert editorial privileges that she had not allowed any other previous collaborator. The passage also prompted scholars to have a closer look at the authorship, style, and content of these visionary texts.*

A twelfth-century authority like the Cistercian abbot **Bernard of Clairvaux (1090–1153)**, one of the most prolific and influential medieval authors, **is known to have been surrounded by a team of secretaries**. For his sermons and letters in particular, he was assisted by a number of collaborators to whom he could dictate his messages or who were asked to produce texts in accordance with his own views. **Some of his collaborators were even trained in imitating his writing style, thus facilitating Bernard's work of final editing or correcting.**

For the present study, Brepols Publishers generously provided a **digital corpus containing the nearly complete works of Hildegard, Guibert, and Bernard of Clairvaux.** We obtained these texts in raw format, corresponding to the way they are included in the Brepols electronic Library of Latin Texts, on the basis of modern critical editions.

Fortunately, these editions are all based on manuscripts that were compiled under the supervision of the original authors or at least in their close vicinity, so that we do not have to worry about major scribal interventions.

## Results

Very tight grouping of writing samples by the different authors in the corpus. We have a clear clustering of samples based on their authorship, so we have a good model, a good fit of the use of the function words by these authors.

If we would be confronted with a new, anonymous sample, we would be able to attribute it to the correct, corresponding author.

With the **oppose() function**, users can contrast two sets of documents and extract the most characteristic features in both sets of texts. The most discriminative features can be visualized and fed into other components of the package as part of a pipeline. Several metrics are implemented that can select features which display a statistically significant difference in distributions between both sets. Craig's Zeta, for instance, is an extension of the Zeta metric originally proposed by Burrows (Burrows, 2007), which remains a popular choice in the stylometric community to select discriminative stylometric features in binary classification settings (Craig and Kinney, 2009).

Named after its creator, Hugh Craig, Zeta is a technique designed to differentiate between the writing styles of different authors.

One way would be to look for patterns in the text, such as the words each writer uses. Some words may appear frequently in your stories, and others may appear often in your friend's stories. But there are also words that are common to both of you.

Craig's Zeta method involves identifying the words that are NOT common between two authors. Instead of looking at the most frequent words (which are often common words like "the", "and", "is", etc.), Zeta looks at words that are used moderately frequently by one author but rarely or never by another.

For instance, suppose you use the word "whisper" moderately frequently in your writing, while your friend rarely uses it. This word then becomes a marker of your writing style.

The 'Zeta score' for a word is a measure of how strongly the word is associated with one author compared to another. A high Zeta score indicates a word that is used distinctly by one author and not the other.