A Statistical Exploration of Text Partition Into Constituents: The Case of the Priestly Source in the Books of Genesis and Exodus

Gideon Yoffe
gideon.yoffe@mail.huji.ac.il

Axel Bühler
axel.buhler@unige.ch

Nachum Dershowitz
nachumd@tauex.tau.ac.il

Israel Finkelstein
fink2@tauex.tau.ac.il

Eli Piasetzky
eip@tauphy.tau.ac.il

Thomas Römer
thomas.romer@college-de-france.fr

Barak Sober
barak.sober@mail.huji.ac.il

Abstract

We present a pipeline for a statistical textual exploration, offering a **stylometry-based explanation and statistical validation of a hypothesized partition of a text**. Given a parameterization of the text, our pipeline: (**1**) detects literary features yielding the optimal overlap between the hypothesized and unsupervised partitions, (**2**) performs a hypothesis-testing analysis to quantify the statistical significance of the optimal overlap, while conserving implicit correlations between units of text that are more likely to be grouped, and (**3**) extracts and quantifies the importance of features most responsible for the classification, estimates their statistical stability and cluster-wise abundance.

We apply our pipeline to the first two books in the Bible, where one stylistic component stands out in the eyes of biblical scholars, namely, the Priestly component. We identify and explore statistically significant stylistic differences between the Priestly and non-Priestly components.

1 Introduction

It is agreed among scholars that the extant version(s) of the Hebrew Bible is a result of various editorial actions (additions, redaction, and more). As such, it may be viewed as a literary patchwork quilt – whose patches differ, for example, by genre, date, and origin – and the distinction between biblical texts as well as their relation to other ancient texts from the Near East is at the heart of biblical scholarship, theology, ancient Israel studies, and philology. Many debates in this field have been ongoing for decades, if not centuries, with no verdict (e.g., the division of the biblical text into its "original" constituents; (Gunkel, 2006; von Rad, 2001; Wellhausen et al., 2009)). While several paradigms gained prevalence throughout the evolution of this discipline (Wellhausen et al., 2009; von Rad, 1972; Zakovitch, 1980), the jury is still out on many related hypotheses, such that these paradigms are prone to drastic (and occasionally abrupt) changes over time (e.g., Nicholson, 2002).

When scrutinized, the interrelation between various literary features within the text may shed light on its historical context. From it, one may infer the number of authors, time(s) and place(s) of composition, and even the geopolitical, social, and theological setting(s) (e.g., Wellhausen et al., 2009; von Rad, 1972; Knohl, 2010; Römer, 2015).

Such scrutiny thus serves a double purpose: (**1**) to disambiguate and identify features in the text that are insightful of the lexical sources of the partition (e.g., Givón, 1991; Yosef, 2018; van Peursen, 2019); (**2**) to attempt to trace these features to the context of the text's composition (e.g., Koppel et al., 2011; Pat-El, 2021).

Works employing computational methods in text stylometry – a statistical analysis of differences in literary, lexical, grammatical or orthographic style between genres or authors (Holmes, 1998) – have been introduced several decades ago (Tweedie et al., 1996; Koppel et al., 2002; Juola et al., 2006; Koppel et al., 2009; Stamatatos, 2009), with biblical exegesis spurring very early attempts at computerized authorship-identification tasks (Radday, 1970; Radday and Shore, 1985). Since then, these methods have proved useful also in investigating ancient (e.g., Kestemont et al., 2016; Verma, 2017; Kabala, 2020), and biblical texts as well (Koppel et al., 2011; Dershowitz et al., 2014; Roorda, 2015; van Peursen, 2019), albeit to a humbler extent. Finally, statistical-learning-based research, which makes headway in an impressively diverse span of disciplines, is taking its first steps in the context of ancient scripts (Murai, 2013; Faigenbaum-Golovin et al., 2016; Dhali et al., 2017; Popović et al., 2021; Faigenbaum-Golovin et al., 2020). In the biblical context, Dershowitz et al. (2014, 2015) addressed reproducing hypothesized partitions of various biblical corpora with a computerized approach as well, using features such as orthographic

Findings of the Association for Computational Linguistics: ACL 2023, pages 1918–1940
July 9-14, 2023 ©2023 Association for Computational Linguistics

differences and synonyms. In the first work – the Cochran-Mantel-Haenszel (CMH) test was applied as a means of hypothesis testing, with the null hypothesis that the synonym features are drawn from the same distribution. While descriptive statistics were successfully applied to various classification and attribution tasks of ancient texts (see above), uncertainty quantification has been insufficiently explored in NLP-related context (Dror et al., 2018), and in particular – in that of text stylometry.

In this work, we introduce a novel exploratory text stylometry pipeline, with which we: (**1**) find a combination of textual parameters that optimizes the agreement between the hypothesized and unsupervised partitions, (**2**) test the statistical significance of the overlap, (**3**) extract features that are important to the classification, the proportion of their importance, and their relative importance in each cluster. Each stage of this analysis was cross-validated (i.e., applied on many randomly-chosen sub-samples of the corpus rather than applied once on the entire corpus) and tested for statistical stability.

To perform (**2**) in a meaningful way for textual analysis, we had to overcome the fact that label-permutation tests do not consider correlations between units of text, which affect their likelihood of being clustered together. This results in unrealistically optimistic p-values, as, in fact, the hypothesized partition *does* implicitly consider such correlations (by, e.g., grouping texts of a similar genre, subject, etc.). We overcome this by introducing a cyclic label-shift test which preserves the structure of the hypothesized partition, thus conserving the implicit correlations therein. Furthermore, we identify literary features that are *responsible* for the clustering, as opposed to intra-cluster feature selection techniques (e.g., Hruschka and Covoes, 2005; Cai et al., 2010; Zhu et al., 2015), which seek to detect significant features within each cluster. This is also a novel approach to text stylometry.

With this pipeline, we examined the hypothetical distinction between texts of Priestly (P) and non-Priestly (nonP) origin in the books of Genesis and Exodus. The Priestly source is the most agreed-upon constituent underlying the Pentateuch (i.e., Torah). The consensus over which texts are associated with P (mainly through semantic analysis) stems from the stylistic and theological distinction from other texts in the Pentateuch, streamlined across texts associated therewith (e.g., Holzinger, 1893; Knohl, 2007; Römer, 2014; Faust, 2019). Therefore, the distinction between P and nonP texts is considered a benchmark in biblical exegesis.

2 Methodology

2.1 Data – Digital Biblical Corpora

We use two digital corpora of the Masoretic variant of the Hebrew Bible in (biblical) Hebrew: (**1**) a version of the Leningrad codex, made freely available by STEPBible.[1] This dataset comes parsed with full morphological and semantic tags for all words, prefixes, and suffixes. From this dataset, we utilize the grammatical representation of the text through phrase-dependent part-of-speech tags (POS). (**2**) A digitally parsed version of the Biblia Hebraica Stuttgartensia (Roorda, 2018) (hereafter BHSA).[2] In the BHSA database, we consider the lexematic (i.e., words reduced to lexemes) and grammatical representation of the text through POS. The difference between the two POS-wise representations of the text is that (**1**) encodes additional morphological information within tags, resulting in several hundreds of unique tags, whereas (**2**) assigns one out of 14 more "basic" grammatical tags[3] to each word. We refer to the POS-wise representation of (**1**) and (**2**) as "high-res POS" and "low-res POS", respectively.

2.2 Manual Annotation of Partition

From biblical scholars, we receive verse-wise labeling, assigning each verse in the books of Genesis (1533 verses) and Exodus (1213 verses) to one of two categories: P or nonP, made available online[4]. Hereafter, we refer to this labeling as "scholarly labeling".

While the dating of P texts in the Pentateuch remains an open, heavily-debated question (e.g., Haran, 1981; Hurvitz, 1988; Giuntoli and Schmid, 2015), there exists a surprising agreement amongst biblical scholars regarding what verses are affiliated to P (e.g., Knohl, 2007; Römer, 2014; Faust, 2019), amounting to an agreement of 96.5% and 97.3% between various biblical scholars for the books of Genesis and Exodus, respectively. We describe the computation of these estimates in Appendix A.

[1] https://github.com/STEPBible/STEPBible-Data

[2] https://etcbc.github.io/bhsa

[3] https://etcbc.github.io/bhsa/features/pdp/

[4] https://github.com/YoffeG/PnonP

2.3 Text Parameterization

The underlying assumption in this work is that a significant stylistic difference between two texts of a roughly-similar genre (or, indeed, any number of distinct texts) should manifest in simple observables in NLP, such as the utilization of vocabulary (i.e., distribution of words) and grammatical structure.

We consider three parameters whose different combinations emphasize different properties-, and therefore yield different classifications- of the text:

- **Lexemes, low-, and high-res POS**: we consider three representations of the text: words in lexematic form and low- and high-res POS (see §2.1). This parameter tests the ability to classify the text based on vocabulary or grammatical structure.
- **n-gram size**: we consider sequences of consecutive lexemes/POS of different lengths (i.e., n-grams). Different sizes of n-grams may be reminiscent of different qualities of the text (e.g., Suen, 1979; Cavnar and Trenkle, 1994; Ahmed et al., 2004; Stamatatos, 2013). For example, a distinction based on a larger n-gram may indicate a semantic difference between texts or the use of longer grammatical modules in the case of POS (e.g., parallelisms in the books of Psalms and Proverbs (Berlin, 1979)). In contrast, a distinction made based on shorter n-grams is indicative of more embedded differences in the use of language (e.g., Wright and Chin, 2014; Litvinova et al., 2015). That said, both these examples indicate that a "false positive" distinction can be made where there is a difference in genre (e.g., Feldman et al., 2009; Tang and Cao, 2015). This degeneracy requires careful analysis of the resulting clustering or inclusion of genre-specific texts only for the clustering phase.
- **Verse-wise running-window width**: biblical verses have an average length of 25 words. Hence, a single verse may not contain sufficient context that can be robustly classified. This is especially important since our classification is based on statistical properties of features in the text (see §2.4). Therefore, we define a running window parameter, which concatenates consecutive verses into a single super-verse (i.e., a running window of k would turn the ith verse to the sequence of the $i - k : i + k$ verses) to provide additional context.

2.4 Text Embedding

We consider individual verses, or sequences of verses, as the atomic constituents of the text (see §2.3). We use tf-idf (term frequency divided by document frequency) to encode each verse, assigning a relevance score to each feature therein (Aizawa, 2003). Works such as Fabien et al. (2020); Marcińczuk et al. (2021) demonstrate that in the absence of a pre-trained neural language model, tf-idf provides an appropriate and often optimal embedding method in tasks of unsupervised classification of texts. For a single combination of an n-gram size and running-window width (using either lexemes, low- or high-res POS), the tf-idf embedding yields a single-feature matrix $X \in \mathbb{R}^{n \times d}$, where n is the number of verses and d is the number of unique n-grams of rank n.

It is important to note that this work aims to set a benchmark for future endeavors using strictly traditional machinery throughout our analysis. To ensure collaboration with biblical scholars, our methodology allows for full interpretability of the exploration process (see §D.4). This has threefold importance: (**1**) the field of text stylometry, especially that of ancient Hebrew texts, has hitherto been explored statistically and computationally to a limited extent, such that even when utilizing conservative text-embeddings, such as in this work, considerable insight can be gained concerning both the quality of the analysis and the philological question. (**2**) Obtaining benchmark results using traditional embeddings is a pre-condition for implementing more sophisticated yet convoluted embeddings, such as pre-trained language models (e.g. Shmidman et al., 2022) or self-trained/calibrated language models (Wald et al., 2021), which we intend to apply in future works. (**3**) Finally, the interdisciplinary nature of this work and our desire to contribute to the field of biblical exegesis (and traditional philology in general) requires our results to be predominantly interpretable, such that they can be subjected to complementary analysis by scholars from the opposite side of the interdisciplinary divide (Piotrowski, 2012).

2.5 Clustering

We choose the k-means algorithm as our clustering tool of choice (Hastie et al., 2009, Ch. 13.2.1) and hardwire the number of clusters to two, according to the hypothesized P/nonP partition. The justification for our choice of this algorithm is its simple loss-optimization procedure, which is vital to our feature importance analysis and is discussed in detail in §2.8.

We use the balanced accuracy (BA) score (Sokolova et al., 2006) for our overlap statistic, a standard measure designed to address asymmetries between cluster sizes.

Due to the stochastic nature of k-means (Bottou, 2004), every time it is used in this work – it is run 50 times – (with different random initialization) – and the result yielding the smallest k-means loss is chosen.

2.6 Optimizing for Overlap

We perform a 2D grid search over a pre-determined range of n-gram sizes and running-window widths to find the parameters combination yielding the optimal overlap for low- or high-res POS lexemes. We test the statistical stability of each combination of these parameters (i.e., to ensure that the overlap reached by each combination is statistically significant) by cross-validating the 2D grid search over some number of randomly-chosen sub-sets of verses, from which we derive the average overlap value for each combination of parameters and the standard deviation thereof. We describe the optimization process in detail in Appendix B.

2.7 Hypothesis Testing and Validating Results

Through hypothesis testing, we establish the statistical significance of the achieved optimal overlap value between the unsupervised and hypothesized partitions. To derive a p-value from some empirical null distribution, we consider the assumption that the hypothesized partition, manifested in the scholarly labeling, exhibits a specific formulaic partition of chunks of the text. A formulaic partition, in turn, suggests that verses within each of the P (nonP) blocks are correlated – a fact that the standard label-permutation test is intrinsically agnostic of, as it permutes labels without considering potential correlations between verses (Fig. 1 left). Thus, the null distribution synthesized through a series of permutations would represent an overly-optimistic scenario that does not correspond to any conceivable scenario in text stylometry.

To remedy this, we devise a more prohibitive statistical test. Instead of permuting the labels to have an arbitrary order, we perform a cyclic shift test of the scholarly labeling (Fig. 1 right). This procedure retains the scholarly labels' hypothesized *structure* but shifts them across different verses. We perform as many cyclic shifts as there are labels (i.e., number of verses) in each book by skips of twice the largest running-window width considered in the optimization procedure. For each shift, we perform the parameter optimization procedure (see §2.6), where we now have the shifted scholarly labels instead of the original ("un-shifted") ones. Thus, we generate a distribution of our statistic under the null hypothesis, from which we derive a p-value.

In Fig. 1, we present an intuitive scheme where we demonstrate our rationale concerning the hypothesis-testing procedure in text stylometry.

2.8 Feature Importance and Interpretability of Classification

Given a k-means labeling produced for the text, which was embedded according to some combination of parameters (§2.3), we wish to quantify the importance of individual n-grams to the classification, the proportion of their importance, and associate to which cluster they are characteristic of.

Consider the loss function of the k-means algorithm:

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} |S_i| \cdot var(S_i),$$

where $k = 2$ is the number of desired clusters, S is the group of all potential sets of verses, split into k clusters, $|S_i|$ is the size of the ith cluster (i.e., number of verses therein) and $var(S_i)$ is the variance of the ith cluster. That is, the k-means aims to minimize intra-class variance. Equivalently, we could optimize for the *maximization* of the *inter-cluster* variance (i.e., the variance between clusters), given by

$$\underset{S}{\operatorname{argmax}} \sum_{i \in S_1} \sum_{j \in S_2} \|x_i - x_j\|^2. \quad (1)$$

Let $D \in \mathbb{R}^{|S_1| \cdot |S_2| \times d}$ denote the matrix of inter-cluster differences whose columns are D_ℓ, which for $i \in \{1, \dots, |S_1| - 1\}$, $j \in \{1, \dots |S_2|\}$ and $\ell = i \cdot |S_2| + j$ are defined by

$$(D)_\ell = x_i - x_j.$$

Then, applying PCA to D would yield the axis across which Eq. (2.8) is optimized as the first *principal component*. This component represents the axis of maximized variance, and each feature's contribution is given by its corresponding loading. Finally, it can be shown that this principal axis could also be computed by subtracting the centroids of the two clusters. Therefore, to leverage this observation and extract the features' importance, we perform the following procedure:

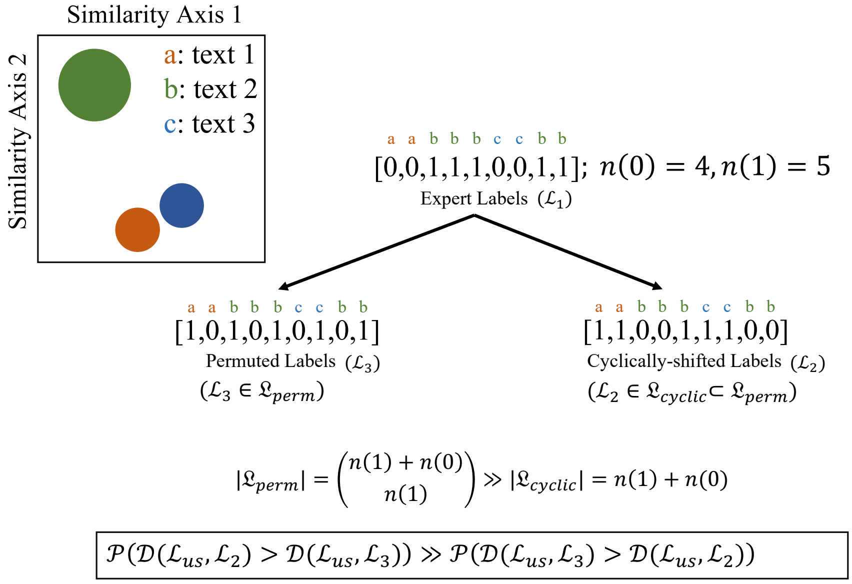

Similarity Axis 1
Similarity Axis 2
a: text 1
b: text 2
c: text 3
a a b b b c c b b
$[0,0,1,1,1,0,0,1,1];\ n(0) = 4, n(1) = 5$
Expert Labels $(\mathcal{L}_1)$
a a b b b c c b b
$[1,0,1,0,1,0,1,0,1]$
Permuted Labels $(\mathcal{L}_3)$
$(\mathcal{L}_3 \in \mathfrak{L}_{perm})$
a a b b b c c b b
$[1,1,0,0,1,1,1,0,0]$
Cyclically-shifted Labels $(\mathcal{L}_2)$
$(\mathcal{L}_2 \in \mathfrak{L}_{cyclic} \subset \mathfrak{L}_{perm})$
$|\mathfrak{L}_{perm}| = \binom{n(1) + n(0)}{n(1)} \gg |\mathfrak{L}_{cyclic}| = n(1) + n(0)$
$\boxed{\mathcal{P}(\mathcal{D}(\mathcal{L}_{us}, \mathcal{L}_2) > \mathcal{D}(\mathcal{L}_{us}, \mathcal{L}_3)) \gg \mathcal{P}(\mathcal{D}(\mathcal{L}_{us}, \mathcal{L}_3) > \mathcal{D}(\mathcal{L}_{us}, \mathcal{L}_2))}$

Figure 1: Hypothesis-testing rationale: consider a case of three distinct texts, divided into nine smaller units, whose relative similarity is plotted in the top left diagram. These units of texts are distinguished into two classes by an expert and are labeled as either 0 or 1 ($\mathcal{L}_1$; where $n(0), n(1)$ are the sums of units of text labeled as 0 and 1, respectively). Here, the correlation between adjacent units of text (which is unknown a-priori) is implicitly taken into account, to some degree, by the expert. Consider now the group of all possible permutations ($\mathfrak{L}_{perm}$) of the expert and the group containing all the possible cyclic shifts thereof ($\mathfrak{L}_{cyclic}$). The latter is a sub-group of – but is much smaller than – the first ($\mathfrak{L}_{cyclic} \subset \mathfrak{L}_{perm}$). For sufficiently long texts, we argue that the probability ($\mathcal{P}$) of randomly permuting the expert labeling such that it would capture some intrinsic correlations between the units of text ($\mathcal{L}_3$) and would yield a substantial overlap (expressed in the figure as the function $\mathcal{D}$, which in our case is the BA score) with an unsupervised clustering labeling ($\mathcal{L}_{us}$) of the text – or at least equivalent to that of the expert labeling – is considerably lower than that of a cyclically-shifted expert labeling ($\mathcal{L}_2$), which, up to a shift, assumes the implicit correlations between units of text manifested in $\mathcal{L}_1$. These, in turn, are more likely to yield equivalent overlap values to $\mathcal{L}_1$. A cyclic-shift test is, therefore, a more prohibitive hypothesis-testing routine than the naive permutation test.

- Compute the *principal separating axis* of the two clusters by subtracting the centroid of S_1 from that of S_2.
- The contribution of each n-gram feature to the cluster separation is given by its respective loading in the principal separating axis.
- Since tf-idf assigns strictly non-negative scores to each feature, the signs of the principal separating axis' loadings allow us to associate the importance of each n-gram to a specific cluster (see Appendix C).
- To determine the stability of the importance of n-grams across multiple sub-samples of each book, we perform a cross-validation routine (similar to the one performed in §2.6). Explicitly, we perform all the steps listed above for some number of simulations (i.e., randomly sample a sub-set of verses and extract the importance vector) and compute the mean and standard deviation of all simulations.
- Finally, we compute the relative uniqueness of each important feature w.r.t. both clusters – assigning it a score between $-1 : 1$. A score of 1 (-1) indicates that the feature is solely abundant in its associated (opposite) cluster. Thus, a feature's abundance nearer zero indicates that its contribution to the clustering is rather in its *combination* with other features than a standalone indicator.

3 Results

We apply a cross-validated optimization analysis to the three representations of the book (§2.6) by performing 2D grid searches for 20 randomly chosen sub-sets of 250 verses for each representation. We perform a cross-validated cyclic hypothesis-testing analysis for the optimal overlap value (§2.7) using five randomly-chosen sub-sets of verses, similarly to the above – and derive a p-value. Finally, we perform a cross-validated feature importance analysis for every representation (§2.8), over 100 randomly-chosen sub-sets of verses, similar to the above.

In Table 1, we list the cross-validated results of each representation for both books and the derived p-value. In Fig. 2, we visualize results for all stages of our analysis applied to the lexematic representation of the book of Genesis. In Appendix D, we plot the results for all stages of our analyses for both books and discuss them in detail. Appendix E presents a detailed biblical-exegetical analysis of our results and an expert's evaluation of our approach.

3.1 Understanding the Discrepancy in Results Between the Books of Genesis and Exodus

In Table 1, we list our optimal overlap values for all textual representations of both books. Notice that there exists a statistically-significant discrepancy between the optimal overlap values between both books that is as follows:

For Exodus, all three representations reach the same optimal overlap of roughly 88%. In contrast, in Genesis – there is a difference between the achieved optimal overlap values for the lexematic and low-res POS representations on the one hand (73%) and high-res POS representation on the other (65%).

When examining the verses belonging to each cluster when classified with the optimal parameter combination, it is evident that most of the overlapping P-associated verses in Exodus are grouped in two blocks of P-associated of text, spanning 243 and 214 verses. These make considerable outliers in the size distribution of P-associated blocks in both books (Fig. 3), which may be related to the observed discrepancy.

This discrepancy begs two questions:

1. Are the linguistic differences between P/nonP – which may be captured by our analysis – that are attenuated for shorter sequences of P texts?
2. Does the high overlap in the book of Exodus arise due to an implicit sensitivity to the generic/semantic uniqueness of the two big P-associated blocks rather than a global stylistic difference between P/nonP?

To examine this, we perform the following experiment: each time, we remove a different P-associated block (1st largest, 2nd largest, 3rd largest, and the 1st + 2nd largest) from the text and perform a cross-validated optimization analysis with low- and high-res POS (see §2.6). We then compare the resulting optimal overlap values of each time. We plot the results of this experiment in Fig. 4.

We find that, as expected, the optimal overlap drops as a function of the size of the removed P-associated text. Interestingly, the optimal overlap *increases* when a smaller block is removed. Additionally, unlike in the case of Genesis, the low-res POS representation doesn't lead to increased optimal overlap relative to the high-res POS representation. This suggests that: (**1**) the fluctuation of the optimal overlap indicates that our pipeline is sensitive to some semantic field associated with the two large P blocks rather than to a global stylistic difference between P/nonP. (**2**) In cases of more sporadically-distributed texts that are stylistically different from the text in which they are embedded – one representation of the text is not systematically preferable to others.

4 Conclusions

We examined the hypothetical distinction between texts of priestly (P) and non-priestly (nonP) origin in the books of Genesis and Exodus, which we explored with a novel unsupervised pipeline for text stylometry. We sought a combination of a running-window width (i.e., the number of consecutive verses to consider as a single unit of text) and n-gram size of lexemes, low- or high-resolution phrase-dependent parts-of-speech that optimized the overlap between the unsupervised and hypothesized partitions. We established the statistical significance of our results using a cyclic-shift test, which we show to be more adequate for text stylometry problems than a naive permutation test. Finally, we extracted n-grams that contribute the most to the classification, their respective proportions, statistical robustness, and correlation to other features. We achieve optimal, statistically significant overlap values of 73% and 90% for the books of Genesis and Exodus, respectively.

We find the discrepancy in optimal overlap values between the two books to stem from two factors: (**1**) A more sporadic distribution of P texts in Genesis, as opposed to a more formulaic one in Exodus. (**2**) The sensitivity of our pipeline to a distinct semantic field manifested in two large P blocks in Exodus, comprising the majority of the P-associated text therein.

Book	Opt. overlap (lexemes)	Opt. overlap (low-res POS)	Opt. overlap (high-res POS)	p-value
Genesis	72.95±6.45% (*rw*: 4, n: 1)	65.03±5.64% (*rw*: 14, n: 1)	73.96±2.91% (*rw*: 4, n: 1)	0.08 (low-res POS)
Exodus	89.23±2.53% (*rw*: 8, n: 2)	88.63±1.96% (*rw*: 9, n: 4)	86.53±2.91% (*rw*: 6, n: 2)	0.06 (high-res POS)

Table 1: Cross-validated optimization and hypothesis testing results: for each representation, we list the optimal overlap value, its respective uncertainty, and combination of parameters (*rw* for running-window width and n for n-gram size).

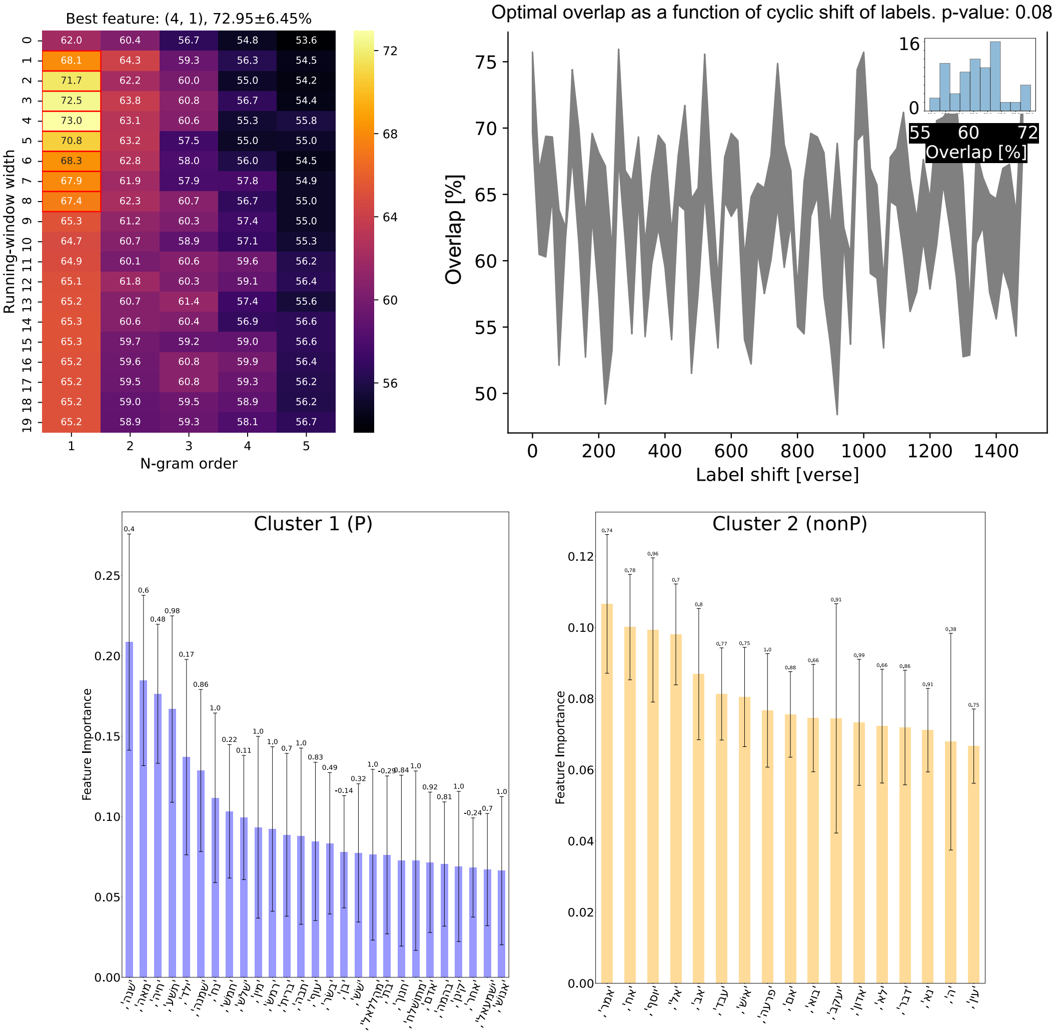


Figure 2: A statistical exploration of the hypothesized partition of the books of Genesis and Exodus into Priestly and non-Priestly constituents: results for the book of Genesis (lexemes representation). **Upper Left Panel – Optimization**: Cross-validated grid-search over verse running-window widths and n-gram sizes to identify the combination yielding an optimal overlap. The combination, value, and uncertainty of the optimal overlap are plotted on top of the panel, and all combinations whose overlap value is within 1σ of the optimal overlap value are marked in red cells. **Upper Right Panel – Hypothesis Testing**: Cross-validated hypothesis testing, where we simulate the null hypothesis distribution through a cyclic shift of the hypothesized P/nonP labeling. The p-value is measured with respect to the optimal overlap value minus its standard deviation. **Bottom Panels – feature importance Analysis**: Cross-validated important feature analysis (running-window 4 and n-gram size 1) and their statistical stability, displaying features bearing 75% of the explained variance. Features in the left and right panels are important to the P-associated and nonP-associated clusters, respectively. The small numbers above each error bar indicate the cluster-wise abundance of the feature.

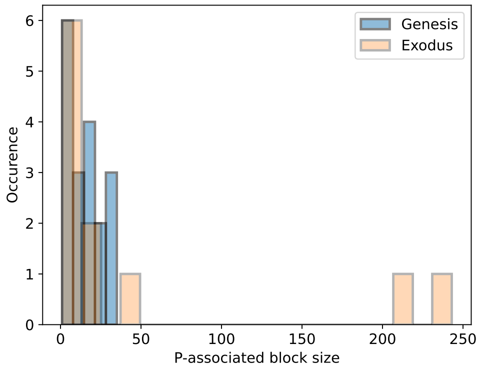


Figure 3: Size distribution histogram of P-associated blocks in the books of Genesis and Exodus.

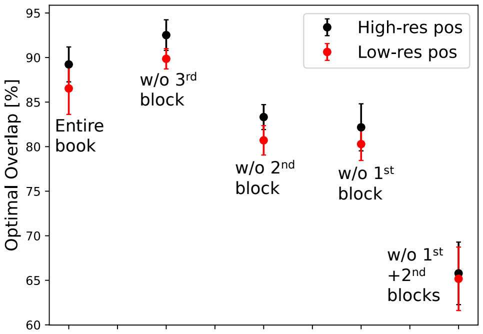


Figure 4: Optimal overlap of a cross-validated optimization analysis of the book of Exodus, after removing different P-associated blocks. **Black dots**: high-res POS. **Red dots**: low-res POS.

Through complementary exegetical and statistical analyses, we show that our methodology differentiates the unique generic style of the Priestly source, characterized by lawgiving, cult instructions, and streamlining a continuous chronological sequence of the story through third-person narration. This observation corroborates and hones the stance of most biblical scholars.

5 Limitations

- The interdisciplinary element – at the heart of this work – mandates that our results be interpretable and relevant to scholars from the opposite side of the methodological divide (i.e., biblical scholars). This, in turn, introduces constraints to our framework – the foremost is choosing appropriate text-embedding techniques. As discussed in §2.4 and §2.8, the ability to extract specific lexical features (i.e., unique n-grams) that are important to the classification, to quantify them, and subject them to complementary philological analysis (see Appendix E) – requires that they be interpretable. This constraint limits the ability to implement state-of-the-art language-model-based embeddings without devising the required framework for their interpretation. Consequently, using traditional embeddings – which encode mostly explicit lexical features (e.g., see §2.4) – limits the complexity of the analyzed textual phenomena and is therefore agnostic of potential signal that is manifested in more complex features.
- In text stylometry questions, especially those related to ancient texts, it is often problematic (and even impossible) to rely on a benchmark training set with which supervised statistical learning can take place. This, in turn, means that supervised learning in such tasks must be implemented with extreme caution so as not to introduce a bias into a supposedly-unbiased analysis. Therefore, implementing supervised learning techniques for such tasks requires a complementary framework that could overcome such potential biases. In light of this, our analysis involves predominantly unsupervised exploration of the text, given different parameterizations.
- Our ability to draw insight from exploring the stylistic differences between the hypothesized distinct texts relies heavily on observing significant overlap between the hypothesized and unsupervised partitions. Without it, the ability to discern the similarity between the results of our pipeline is greatly obscured, as the pipeline remains essentially agnostic of the hypothesized partition. Such a scenario either deems the parameterization irrelevant to the hypothesized partition or disproves the hypothesized partition. Breaking the degeneracy between these two possibilities may entail considerable additional analysis.

6 Acknowledgements

We thank Dr. Rotem Dror for her kind assistance and contribution to the editing process and Ziv Ben-Aharon for providing technical guidance in operating the HUJI cluster.

References

Bashir Ahmed, Sung-Hyuk Cha, and Charles Tappert. 2004. Language identification from text using n-gram based cumulative frequency addition. In

Proceedings of Student/Faculty Research Day, volume 12. CSIS, Pace University.

Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65.

Adele Berlin. 1979. Grammatical aspects of biblical parallelism. *Hebrew Union College Annual*, 50:17–43.

Léon Bottou. 2004. Stochastic learning. In *Summer School on Machine Learning*, pages 146–168. Springer.

Deng Cai, Chiyuan Zhang, and Xiaofei He. 2010. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 333–342.

William B. Cavnar and John M. Trenkle. 1994. N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, volume 161175. Citeseer.

Idan Dershowitz, Navot Akiva, Moshe Koppel, and Nachum Dershowitz. 2015. Computerized source criticism of biblical texts. *Journal of Biblical Literature*, 134(2):253–271.

Idan Dershowitz, Nachum Dershowitz, Tomer Hasid, and Amnon Ta-Shma. 2014. Orthography and biblical criticism. In *Proceedings of Digital Humanities (DH 2014, Lausanne, Switzerland)*, pages 451–453.

Maruf A. Dhali, Sheng He, Mladen Popović, Eibert Tigchelaar, and Lambert Schomaker. 2017. A digital palaeographic approach towards writer identification in the Dead Sea Scrolls. In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods*. SCITEPRESS - Science and Technology Publications.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

Maël Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. Bertaa: Bert fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137.

Shira Faigenbaum-Golovin, Arie Shaus, Barak Sober, David Levin, Nadav Na'aman, Benjamin Sass, Eli Turkel, Eli Piasetzky, and Israel Finkelstein. 2016. Algorithmic handwriting analysis of Judah's military correspondence sheds light on composition of biblical texts. *Proceedings of the National Academy of Sciences*, 113(17):4664–4669.

Shira Faigenbaum-Golovin, Arie Shaus, Barak Sober, Eli Turkel, Eli Piasetzky, and Israel Finkelstein. 2020. Algorithmic handwriting analysis of the Samaria inscriptions illuminates bureaucratic apparatus in biblical Israel. *PLOS ONE*, 15(1):e0227452.

Avraham Faust. 2019. The world of P: The material realm of priestly writings. *Vetus Testamentum*, 69(2):173–218.

Sergey Feldman, Marius A. Marin, Mari Ostendorf, and Maya R. Gupta. 2009. Part-of-speech histograms for genre classification of text. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4781–4784. IEEE.

Federico Giuntoli and Konrad Schmid. 2015. *The Post-Priestly Pentateuch: New Perspectives on its Redactional Development and Theological Profiles*. Mohr Siebeck.

Talmy Givón. 1991. The evolution of dependent clause morpho-syntax in Biblical Hebrew. In Elizabeth Closs Traugott and Berund Heine, editors, *Approaches to SGrammaticalization*, volume 2, pages 257–310. John Benjamins.

Hermann Gunkel. 2006. *Creation and Chaos in the Primeval Era and the Eschaton: Religio-Historical Study of Genesis 1 and Revelation 12*. Eerdmans, Grand Rapids, MI. Trans. by K. William Whitney, Jr.; original edition 1895.

Mehahem Haran. 1981. Behind the scenes of history: Determining the date of the priestly source. *Journal of Biblical Literature*, 100(3):321–333.

T. Hastie, R. Tibshirani, and J. H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer.

David I. Holmes. 1998. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–117.

Heinrich Holzinger. 1893. *Einleitung in den Hexateuch*, volume 1. Mohr Siebeck.

Eduardo R. Hruschka and Thiago F. Covoes. 2005. Feature selection for cluster analysis: an approach based on the simplified silhouette criterion. In *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, volume 1, pages 32–38. IEEE.

Avi Hurvitz. 1988. Dating the priestly source in light of the historical study of biblical Hebrew. a century after Wellhausen. *Zeitschrift für die alttestamentliche Wissenschaft*, 100:88–100.

Louis C. Jonker. 2012. Reading the Pentateuch's genealogies after the exile: The Chronicler's usage of Genesis 1–11 in negotiating an all-Israelite identity. *Old Testament Essays*, 25(2):316–333.

Patrick Juola, John Sofko, and Patrick Brennan. 2006. A prototype for authorship attribution studies. *Literary and Linguistic Computing*, 21(2):169–178.

Jakub Kabala. 2020. Computational authorship attribution in medieval Latin corpora: the case of the Monk of Lido (ca. 1101–08) and Gallus Anonymous (ca. 1113–17). *Language Resources and Evaluation*, 54(1):25–56.

Mike Kestemont, Justin Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. 2016. Authenticating the writings of Julius Caesar. *Expert Systems with Applications*, 63:86–96.

Israel Knohl. 2007. *The Sanctuary of Silence: The Priestly Torah and the Holiness School*. Eisenbrauns.

Israel Knohl. 2010. *The Divine Symphony: The Bible's Many Voices*. Jewish Publication Society.

Moshe Koppel, Navot Akiva, Idan Dershowitz, and Nachum Dershowitz. 2011. Unsupervised decomposition of a document into authorial components. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1356–1364, Portland, OR. Association for Computational Linguistics.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

T. A. Litvinova, P. V. Seredin, and O. A. Litvinova. 2015. Using part-of-speech sequences frequencies in a text to predict author personality: a corpus study. *Indian Journal of Science and Technology*, 8:93.

Michał Marcińczuk, Mateusz Gniewkowski, Tomasz Walkowiak, and Marcin Będkowski. 2021. Text document clustering: Wordnet vs. TF-IDF vs. word embeddings. In *Proceedings of the 11th Global Wordnet Conference*, pages 207–214.

Hajime Murai. 2013. Exegetical Science for the Interpretation of the Bible: Algorithms and Software for Quantitative Analysis of Christian Documents. In *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, pages 67–86. Springer International Publishing.

Ernest Nicholson. 2002. *The Pentateuch in the Twentieth Century*. Oxford University Press.

Na'ama Pat-El. 2021. Syntactic Aramaisms as a tool for the internal chronology of Biblical Hebrew. In *Diachrony in Biblical Hebrew*, pages 245–264. Penn State University Press.

Michael Piotrowski. 2012. NLP and digital humanities. In *Natural Language Processing for Historical Texts*, pages 5–10. Springer.

Mladen Popović, Maruf A. Dhali, and Lambert Schomaker. 2021. Artificial intelligence based writer identification generates new evidence for the unknown scribes of the Dead Sea Scrolls exemplified by the Great Isaiah Scroll (1QIsaa). *PLOS ONE*, 16:1–28.

Yehuda T. Radday. 1970. Isaiah and the computer: A preliminary report. *Computers and the Humanities*, 5(2):65–73.

Yehudah T. Radday and Haim Shore. 1985. *Genesis: An authorship study in computer-assisted statistical linguistics*, volume 103 of *Analecta Biblica*. Biblical Institution Press.

Thomas Römer. 2014. From the call of Moses to the parting of the sea: Reflections on the priestly version of the Exodus narrative. In *The Book of Exodus*, pages 121–150. Brill.

Thomas Römer. 2015. *The Invention of God*. Harvard University Press.

Dirk Roorda. 2015. The Hebrew Bible as Data: Laboratory-Sharing-Experiences. *CLARIN in the Low Countries*.

Dirk Roorda. 2018. Coding the Hebrew Bible: Linguistics and literature. *Research Data Journal for the Humanities and Social Sciences*, 3(1):27–41.

Avi Shmidman, Joshua Guedalia, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Eli Handel, and Moshe Koppel. 2022. Introducing berel: Bert embeddings for rabbinic-encoded language. *arXiv preprint arXiv:2208.01875*.

Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. 2006. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian Joint Conference on Artificial Intelligence*, pages 1015–1021. Springer.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Efstathios Stamatatos. 2013. On the robustness of authorship attribution based on character *n*-gram features. *Journal of Law and Policy*, 21(2):421–439.

Ching Y. Suen. 1979. N-gram statistics for natural language understanding and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):164–172.

Xiaoyan Tang and Jing Cao. 2015. Automatic genre classification via n-grams of part-of-speech tags. *Procedia-Social and Behavioral Sciences*, 198:474–478.

Fiona J. Tweedie, Sameer Singh, and David I. Holmes. 1996. Neural network applications in stylometry: The Federalist papers. *Computers and the Humanities*, 30(1):1–10.

Wido van Peursen. 2019. A Computational Approach to Syntactic Diversity in the Hebrew Bible. *Journal of Biblical Text Research*, 44:237–253.

Mayuri Verma. 2017. Lexical Analysis of Religious Texts using Text Mining and Machine Learning Tools. *International Journal of Computer Applications*, 168(8):39–45.

Gerhard von Rad. 1972. *Genesis – A commentary*, 3rd rev. ed. edition. S.C.M. Press, London. Trans. by John H. Marks.

Gerhard von Rad. 2001. *Old Testament Theology: The theology of Israel's historical traditions*, volume 1. Westminster John Knox Press.

Yoav Wald, Amir Feder, Daniel Greenfeld, and Uri Shalit. 2021. On calibration and out-of-domain generalization. *Advances in neural information processing systems*, 34:2215–2227.

Julius Wellhausen, J. Sutherland Black, Allan Menzies, and William Robertson Smith. 2009. *Prolegomena to the History of Israel*. Cambridge University Press.

William R. Wright and David N. Chin. 2014. Personality profiling from text: Introducing part-of-speech n-grams. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 243–253. Springer.

Ofer Yosef. 2018. Determining orthography on the basis of Masoretic notes. In *The Masora on Scripture and Its Methods*, chapter 3, pages 34–48. De Gruyter, Berlin.

Yair Zakovitch. 1980. A study of precise and partial derivations in biblical etymology. *Journal for the Study of the Old Testament*, 5(15):31–50.

Guo-Niu Zhu, Jie Hu, Jin Qi, Jin Ma, and Ying-Hong Peng. 2015. An integrated feature selection and cluster analysis techniques for case-based reasoning. *Engineering Applications of Artificial Intelligence*, 39:14–22.

Appendices

A Estimating Scholarly Consensus of P-associated Texts in Genesis and Exodus

To quantify the consensus amongst biblical scholars concerning the distinction between P/nonP texts in Genesis and Exodus, we consider two sets of labelings of P/nonP: the first is that provided by biblical scholars, and the other is similar labeling used in the work of Dershowitz et al. (2015), who also apply computational methods in an attempt to detect a meaningful dichotomy between P and nonP texts as well, albeit under a different paradigm. In their work, Dershowitz et al. consider P/nonP distinctions of three independent biblical scholars and compile a single "consensus" labeling of verses over the affiliation of which all three scholars agree (1291 verses in Genesis and 1057 verses in Exodus), except for explicitly incriminating P texts (such as genealogical trees that are strongly affiliated with P (e.g., Jonker, 2012)). We find a 96.5% and 97.3% agreement between the two labelings for the books of Genesis and Exodus, respectively.

B Cross-validated Overlap Optimization

The cross-validated overlap optimization is performed as follows:

- For lexemes/POS, we consider a range of n-gram sizes and a range of running-window widths, through combinations of which we optimize the classification overlap. These ranges are determined empirically by observing the overlap values decrease monotonically as the n-gram sizes/running-window widths become too large (see Figs. 6, 5). This produces a 2D matrix, where the (i, j)th entry is the resulting overlap value achieved by the k-means given the combination of the ith running-window width and the jth n-gram size.
- We cross-validate the grid-search process by performing a series of simulations, where in each simulation, we generate a random subset of 250 verses from the given book, containing at least 50 verses of each class (according to the scholarly labeling). Each such simulation produces a 2D overlap matrix, as mentioned above, for the given subset of verses.
- Finally, we average the 2D overlap matrices

of all simulations, and for each entry, we calculate its standard deviation across all simulations, producing a 2D standard deviation matrix. After normalizing the average overlap matrix by the standard deviation matrix, we choose the optimal combination that produces the classification with the optimal overlap.

- We perform this analysis on both lexemes, low- and high-res POS, yielding three averaged overlap matrices, from which optimal parameter combinations and their respective standard deviations (calculated across the cross-validation simulations) can be identified for each feature.

C Feature Importance Analysis

We consider the two optimal overlap clusters of verses as the algorithm assigned them. We calculate a difference matrix D, where from every verse in cluster 1, we subtract every verse in cluster 2, receiving a matrix $D \in \mathbb{R}^{|S_1|\cdot|S_2|\times d}$, where $|S_1|$, $|S_2|$ are the number of verses assigned to cluster 1 and 2, respectively, and d is the dimension of the embedding (i.e., number of unique n-grams in the text). Note that tf-idf embedded texts are non-negative, such that the difference matrix D has positive values for features with a high tf-idf score in cluster 1 and negative values for features with a high tf-idf score in cluster 2. As mentioned in §2.8, the first principal axis of D is equivalent to the difference between the two centroids produced by the k-means. Similarly, this difference vector is a linear combination of all the features (i.e., n-grams), where a numerical value gives the importance of each feature called "loading", ranging from negative to positive values and their importance to the given principal axes is given by their absolute value. Due to the nature of the difference matrix D (see §2.8) – the sign of the loading indicates in which cluster the feature is important. Thus we can assign distinguishing features to the specific cluster of which they, or a combination thereof with other features, are characteristic.

Finally, we seek to determine the stability of the importance of features across multiple sub-samples of each book. We perform this computation as follows: given a parameters combination, we perform all the steps listed above for 100 simulations, where in each simulation, we randomly sample a sub-sample of 250 verses and extract the importance loadings of the features. We then average all simulation-wise loadings and derive the variance thereof to receive a cross-validated vector of feature importance loadings and their respective uncertainties. These are plotted as the error bars in Figs. 2, and similarly in the figures in Appendix C

D Results

D.1 Optimization Results: Genesis

For each representation, we achieve the following optimal overlap values (see Fig. 6): 72.95±6.45% for lexemes, 65.03±5.64% for high-res POS, and 73.96±2.91% for low-res POS. We observe the following:

- Optimal overlap values achieved for lexemes and high-res POS are consistent to within a 1σ, whereas the optimal overlap achieved for low-res POS is higher by $\sim 2\sigma$.
- We find a less consistent and considerably larger spread of optimal parameter combinations for the low- and high-POS-wise embeddings, as opposed to Exodus. While no clear pattern of optimal parameter combinations is observed in any feature, small n-gram sizes are also preferred here. For lexemes, a well-defined range of running-window widths of unigrams is observed to yield optimal overlaps.
- Unlike in the case of Exodus, parameter combinations yielding optimal overlap values of each feature (marked with red cells in Fig. 5) do not exhibit higher consistency across the cross-validation simulations than combinations that yield smaller overlap (i.e., small cross-validation variance, see the right panels in Fig. 5), except for the low-res POS feature.

D.2 Optimization Results: Exodus

For each feature, we achieve the following optimal overlaps (see Fig. 6): 89.23±2.53% for lexemes, 88.63±1.96% for high-res POS, and 86.53±2.91% for low-res POS. We observe the following:

- For all three representations, optimal overlap values are consistent to within 1σ.
- For all three representations, parameter combinations yielding optimal overlap values (marked with red cells in Fig. 6) exhibit high consistency across the cross-validation simulations (i.e., slight cross-validation variance, see the right panels in Fig. 6).

- For lexemes, we find that the range of optimal parameter combinations is concentrated within 1- and 2-grams and is relatively independent of running-window width (i.e., when some optimal running-window width is reached–the overlap values do not change dramatically as it increases).
- For the high-res POS, we find that the range of optimal combinations is restricted to 2-grams, but is also insensitive to the running-window width.
- For low-res POS, we find that larger n-gram sizes, and a wider range thereof, produce the optimal overlap values. Additionally, we observe a dependence between given n-gram sizes and running-window widths to reach a large overlap.

D.3 Hypothesis Testing Results

We present our results of the hypothesis testing through cyclic-shift, described in §2.7. Here, too, we perform a cross-validated test by performing five simulations – each containing a randomly chosen sub-sample of 250 verses (with a mandatory minimum of 50 verses of each class), to which the cyclic-shift analysis is applied. We compute the optimal overlap for every shift and generate a shift-series of optimal overlap values (i.e., the null distribution). We then average across simulations. We then derive the p-value from the synthesized null distribution. The chosen "real optimal overlap", which we use to derive the p-value, is the average optimal overlap at a shift of 0 (i.e., original labeling) minus its standard deviation. For each book, we perform this analysis for the features yielding the optimal overlap; low-res POS for Genesis and high-res POS for Exodus. We plot our results for both books in Fig. 7.

The resulting p-values are 0.08 and 0.06 for the books of Genesis and Exodus, respectively.

D.4 Feature Importance Analysis Results

D.4.1 Feature Importance: Genesis

In Figs. 8-10 we plot feature importance analysis results for the three representations of the book of Genesis.

D.4.2 Feature Importance: Exodus

In Figs. 11-13 we plot feature importance analysis results for the three representations of the book of Exodus.

E Biblical-Exegetical Discussion

Here we perform an exegetical analysis of our results for each book. All data to which this analysis was applied is available online[5].

E.1 Genesis

E.1.1 Semantics

The extraction of the features of P for Genesis overlaps with the work of characterizing the priestly stratum (Holzinger, 1893). Thus, the characteristic use of numbers in P (here, in descending order of importance, the algorithm considered the terms 100, 9, 8, 5, 3, 6, 4, 2, 7 as characteristic of P) appears mainly in the genealogies, e.g., Gen 5 and 11, but also in the use of ordinal numbers to give the months and in the definition of the dimensions of the tabernacle. The term "year" (שׁנה) is used in both dates and P genealogies, and the term "day" (יום) demonstrates a similar calendrical concern. Furthermore, in the genealogies, we find the names of the patriarchs considered to belong to P (נח, מהללאל, קינן, חנוך, מתושלח, אנוש, אדם, שׁת, ישׁמעאל, למך, רעו, פלג, סרוג). The term "son" (בן) appears in genealogies but also in typical P expressions such as "son of X year" (בן שׁנה) to indicate the age of someone, "sons of Israel" (בני ישׂראל), etc. The root ילד (to beget or to give birth) is found in the genealogies in Gen 5; 11; Exod 6 but also in other P narratives of the patriarchs (Gen 16–17; 21; 25; 35; 36; 46; 48) which focus on affiliation. The term "generation" (דור) is also recognized as typically P (Gen 6:9; 9:12; 17:7,9,12; etc.), as well as the term "annals" (תולדות), which serves to introduce a narrative section or a genealogy and. This term structures the narrative and genealogical sections in the book of Genesis (Gen 2:4; 5:1; 6:9; 10:1,32; 11:10,27; 25:12–13,19; 36:1,9; etc.). The terms "fowl" (עוף), "beast/flesh" (בשׂר), "creeping" (רמשׂ), "swarming" (שׁרץ), "living being" (נפשׁ חיה), "cattle" (בהמה), "kind" (מין) are found in the typically P expression "living creatures of every kind: cattle and creeping things and wild animals of the earth of every kind" (Gen 1:24; cf. Gen 1:25-26; 6:7,20; 7:14,23; etc.). These expressions are often associated with "multiplication" (רבה), an essential theme for P that also appears in the blessings of P narratives as in Gen 17; 48; etc. The term "being" (נפשׁ) is also used in P texts to refer to a person, e.g., in Gen 12; 17; 36; 46. As for the term "all" (כל), it

[5] https://github.com/YoffeG/PnonP

is used overwhelmingly in both P and D texts. In P texts (Gen 1:27; 5:2; 6:19), humanity (אדם), in the image (צלם) of God, is conceived in a dichotomy of "male" (זכר) and "female" (נקבה). The root זכר in its second sense, that of "remembering", also plays a role in the P narratives (Gen 8:1; 9:15–16; 19:29; Exod 2:24; 6:5) when God remembers his covenant and intervenes to help humanity or the Israelites. The covenant (ברית; cf. also Gen 9; 17), the sign of which is the circumcision (מול) of the foreskin (ערלה) is correctly characterized as P. According to P, God's covenants are linked to a promise of offspring (זרע; cf. Gen 17; etc.) and valid forever (עולם; Gen 9; 17; 48:4; Exod 12:14; etc.). The term "seed/descendant" (זרע) is also used by P in the creation narratives in Gen 1. The term "between" (בין) is used several times to indicate the parties concerned by the covenant in Gen 9 and Gen 17. The term is also found frequently in Gen 1 in the creation story, where creation is the result of separation "between" (בין) different elements – presenting God as the creator is not typical of a national god whose role primarily guarantees protection, military success, and fertility. The transformation of the God of Israel into a creator God appears only in the exilic or postexilic texts. Thus, the root "to create" (ברא) is rightly associated with P (Gen 1:1-2:4; 5:1-2). The use of divine names is particular in the priestly narratives. "God" (אלהים) is the term used in the origin stories (Gen 1–11), "El Shaddai" for the patriarchs (Gen 12–Exod 6), and finally, "YHWH" from Exodus 6,2-3 on. Here, the algorithm did understand a particular use by P of the term "God" (אלהים). One of the differences with Holzinger's list is the fact that the algorithm considers the terms Noah (נח), flood (מבול), and the ark (תבה) as typically P. This is probably because the flood narrative is much more developed in P than in non-P or because the semantic environment is attached to other P expressions. Nevertheless, all three terms appear in non-P texts as well. The term "daughter" (בת) should be considered P not because of its frequency, which is admittedly somewhat higher in the P narratives of Genesis, but probably because of its semantic environment. Thus, the term appears in the expression "sons and daughters" (בנות ובנים), which is very frequently used in Gen 5; 11. The preposition "after" (אחר) appears in the expression "after you" (אחריך) in the promise to Abraham in Gen 17 or the expression "after his begetting" (הולידו אחרי) in the genealogies in Gen 5; 11. The appearance of the term "to die" (מות) as a characteristic of P is explained by its presence in the genealogies of Gen 5; 11 but also in the succession of each of the generations of the patriarchs. Finally, the terms "water" (מים) and "heaven" (שמים) play a major role in the creation narrative P (Gen 1:1-2:4) and the flood narrative (Gen 6-9*). These two terms also appear in Exodus, where water is mentioned in the account of the duel of the magicians (Exod 7-9*), in the passage of the sea of reeds which is paralleled in the creation of Gen 1 (Exod 14), and as a means of purification during the building of the tabernacle (Exod 29-30; 40). This latter function of water probably builds its symbolism in the other narratives. The term firmament (רקיע) is associated with heaven and appears only in the creation story P of Gen 1 but is of little significance elsewhere.

On the non-P side, terms like "Joseph", "pharaoh", and "Egypt" are non-P features since the story of Joseph is non-P. Similarly, the presence of "Jacob" is explained by an account of only a few verses for this story in P as opposed to several whole chapters for the non-P account of Jacob. The terms "brother" ($16^{P}/179^{nonP}$), "father" ($19^{P}/213^{nonP}$), and "mother" ($4^{P}/33^{nonP}$) as non-P features can be understood by a greater emphasis on family in the original patriarchal accounts, whereas P emphasizes genealogy. The terms "master" (אדון), "slave/servant" (עבד), and "boy/servant" (נער) reflect the hierarchical structures of the household of the wealthy landowners in the narratives of the patriarchs but are of no interest to the priestly editors. Similarly, non-P texts show more interest in livestock, with terms such as camel (גמל), donkey (חמור), or small livestock (צאן) considered non-P features. The dialogues are more present in the non-P stories than in the P stories. Thus the terms that open the direct discourses "speak" (דבר), "say" (אמר), and "tell" (נגד) are considered typical non-P terms as well as the set of Hebrew propositions in direct discourse (מה, עתה, ה, כי, גם, הנה, אל, נא, אם, לא). The term "man" (איש) can be used in many ways: man, husband, human; someone. Its use and expressions using it are significantly more frequent in non-P texts (42P/213nonP). This may be an evolution of the language rather than a deliberate or theological change on the part of P. Finally, for the terms "to enter" (בוא) or "to go" (הלך) to be characteristic of non-P, this may reflect a stronger interest in place, in travel in the original texts probably composed to legitimize sanctuaries or as etiological narratives whereas these aspects are less marked in the P texts.

E.2 Exodus

E.2.1 Semantics

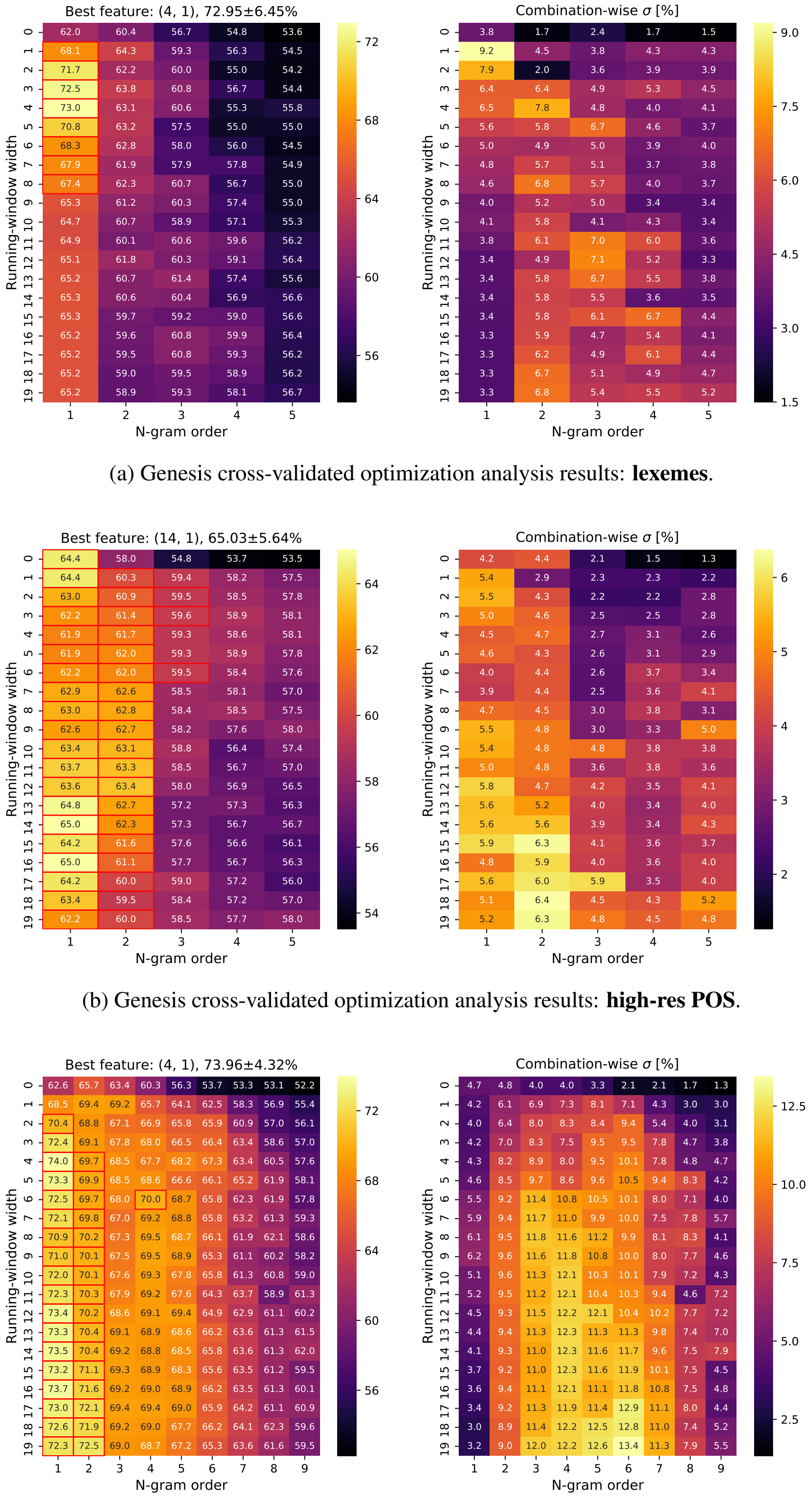
For the P-texts of Exodus, the algorithm has extracted the semantic features of the tabernacle construction in Exod 25-31; 35-40 but does not give features of the P-texts that would be found elsewhere. We find in the features: the different names of the holy place, "the holy one", "the dwelling", "the tent of meeting"; the materials used for the construction, "acacia wood", "pure gold", "bronze", "linen", "blue, purple, crimson yarns", etc.; the spatialization, "around", "outside"; the dimensions, "length", "cubit", "five"; the components, "altar", "curtain", "ark", "utensils", "table", and YHWH's orders to Moses, "you shall make". Thus, the algorithm has a good understanding of the terms specific to the construction of the Tabernacle but is not susceptible to a more general understanding of the characteristics of P in Exodus. The non-P features are more interesting. For example, the use of the word "people" (עם) appears primarily in the non-P texts because the priestly redactors usually preferred to use the term "assembly" (עדה). The word "I" in the long form (אנכי) is considered non-P because the word 'ny, the short form, appears in P texts. The expression "to YHWH" does appear 24 times in non-P texts, e.g., "to cry out to YHWH"/"to speak to YHWH"/"to turn to YHWH", whereas P avoids the expression. This is easily understandable by a desire to give YHWH the initiative in all interactions. In P, it is he who demands, commands, and speaks. There are few dialogues. As for the terms Egypt (מצרים) and Pharaoh (פרעה), they are indeed quantitatively more frequent in the non-Priestly texts of Exodus (respectively 36P/139$^{\text{nonP}}$ et 26P/89$^{\text{nonP}}$) as in Genesis.

E.2.2 Grammar

As we have already seen, non-P texts more often adopt the protagonists' point of view by including dialogues or their thoughts, whereas P texts prefer a third-person narration. One of the consequences is the privileged use of 3rd singular or plural suffixes, unlike non-P texts where 1st singular or 2nd singular suffixes are more often used. Moreover, the massive use of the 3rd person in P texts can also be explained by the presence of pleonasms which use a form with this suffix: עמו, אתו, etc. (Holzinger, 1893). Concerning verbs, the form Qal or Piel, qatal in the 2nd masculine singular, is prevalent in P texts. This is understandable because of P's theology, according to which God orders using the second person, and then the protagonists carry out according to YHWH's order. On the side of the non-P texts, the narrative form, i.e., Qal, wayyiqtol in the 3rd masculine singular, is significant, although these forms are also very present in P texts. Another peculiarity is the use in non-P of "name in the constructed form + place name". P seems to have avoided topical constructions because of reduced interest in localizations. The remaining terms are persistent elements. Further analysis would be needed to understand the relevance of the distinction made by the algorithm.

E.3 Summary

As we can see, the pipeline could extract typical features of priestly texts in Genesis, easily recognizable for a specialist. In addition, other P features have also been found that may be specific to a single narrative, correspond to repeated use of an expression, or be a significant theological theme such as water. On the other hand, the features of non-P texts do not indicate a coherent editorial milieu or style but rather allow us to better distinguish between P texts and non-P texts by particular theological or linguistic features. The data provided by the algorithm allow for the detection of particularities that require an explanation. More detailed investigations than those presented above are necessary to better understand specific instances of the results. For the texts of Exodus, the excessive importance of the chapters devoted to the construction of the Tabernacle (Exod 25-31; 35-40) did not allow us to obtain satisfactory results in the characterization of P, which could indicate an originally independent document. Nevertheless, the characterization of non-P texts is relevant, as well as the results on grammar.

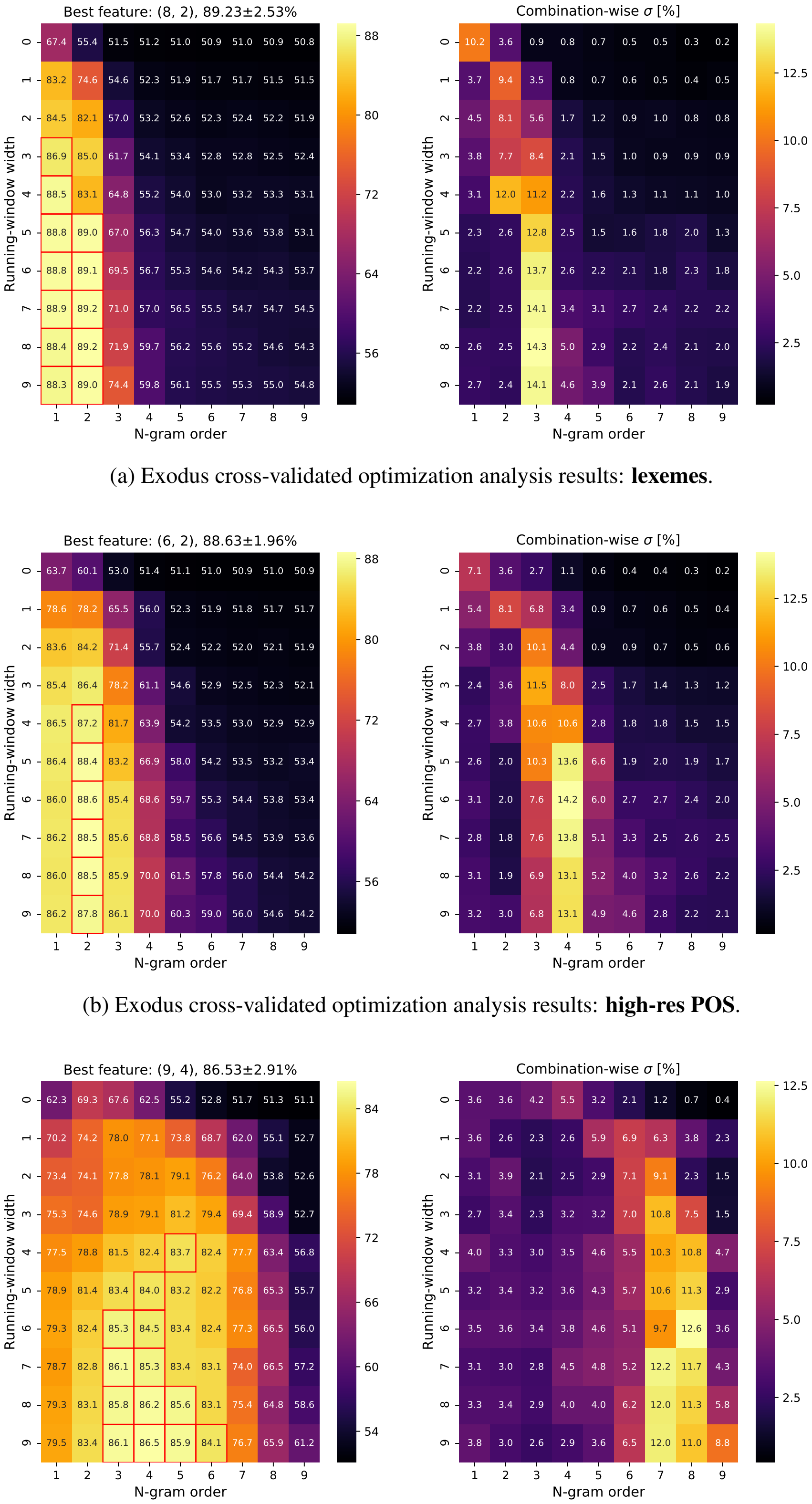


(a) Genesis cross-validated optimization analysis results: **lexemes**.

(b) Genesis cross-validated optimization analysis results: **high-res POS**.

(c) Genesis cross-validated optimization analysis results: **low-res POS**.

Figure 5: Cross-validated optimization analysis results for the book of Genesis, using all three representations: lexemes (**a**), high- and low-res POS (**b**, **c**), respectively, probing a running-window width range of 1–10 and n-gram size range of 1–10 over 20 simulations. **Left Panels**: Averaged combination-wise overlaps. **Right Panels**: Combination-wise standard-deviation (σ) of the overlaps generated with the cross-validation process. For each case, the best-fit feature (running-window width, n-gram size) and its respective σ appear at the top of each left panel. Overlaps of combinations within the 1σ range of the best fit are marked with red cells.

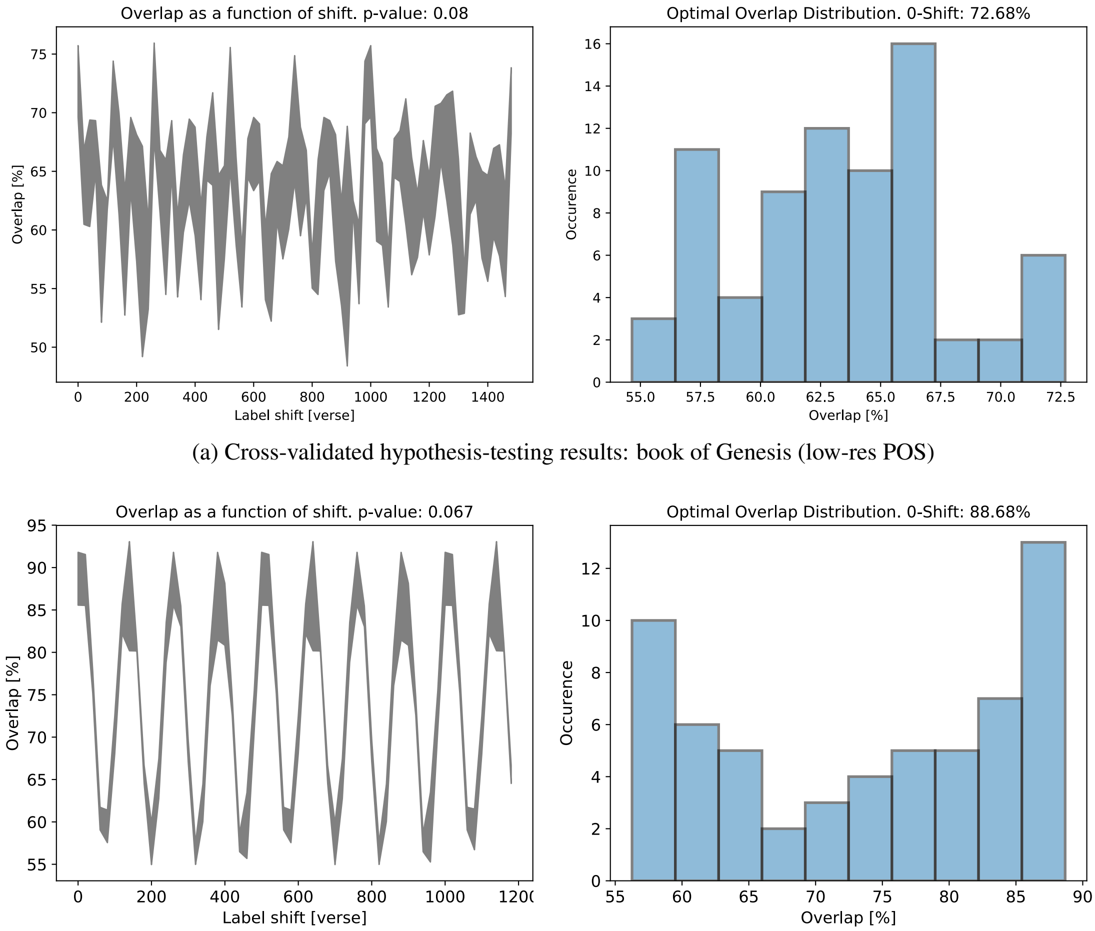


(a) Exodus cross-validated optimization analysis results: **lexemes**.

(b) Exodus cross-validated optimization analysis results: **high-res POS**.

(c) Exodus cross-validated optimization analysis results: **low-res POS**.

Figure 6: Cross-validated optimization analysis results for the book of Exodus, similar to Fig. 5.



(a) Cross-validated hypothesis-testing results: book of Genesis (low-res POS)

(b) Cross-validated hypothesis-testing results: book of Exodus (high-res POS)

Figure 7: Cross-validated hypothesis-testing results, performed as described in §3, for the books of Genesis (sub-figure **a**) and Exodus (sub-figure **b**), respectively. **Left Panels**: Optimal overlap as a function of cyclic label shift. The derived p-value is listed on top of each panel. **Right Panels**: Resulting null distributions of the optimal overlap values. The optimal overlap value of the 0-shift (i.e., scholarly labeling) is listed on each panel.

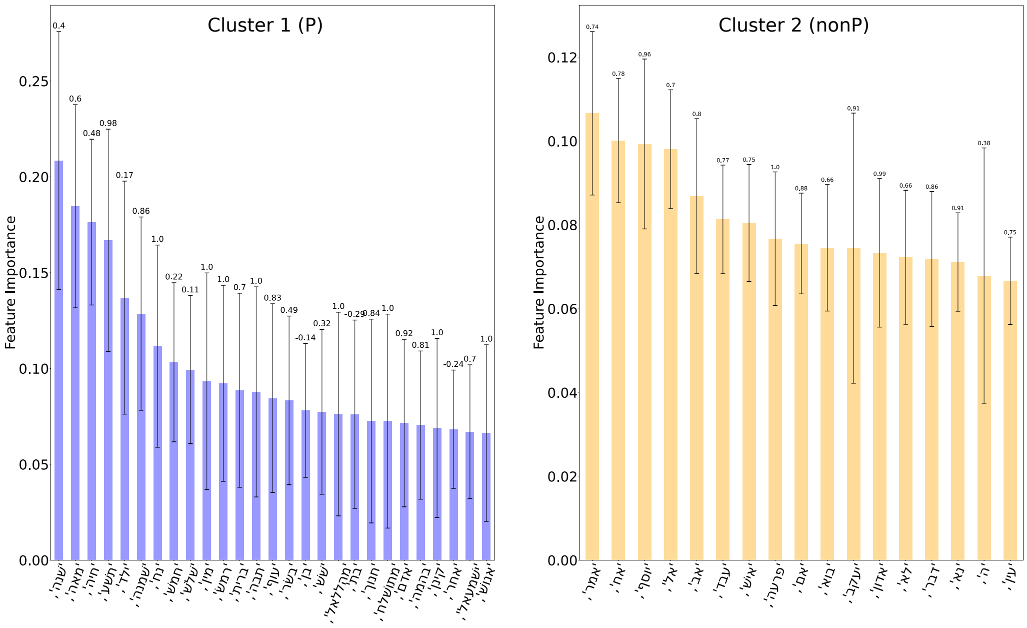


Figure 8: Cross-validated feature importance analysis results for the book of Genesis: **lexemes** (running-window width of 4; n-gram size of 1). **Note**: displaying features carrying 75% of the explained variance.

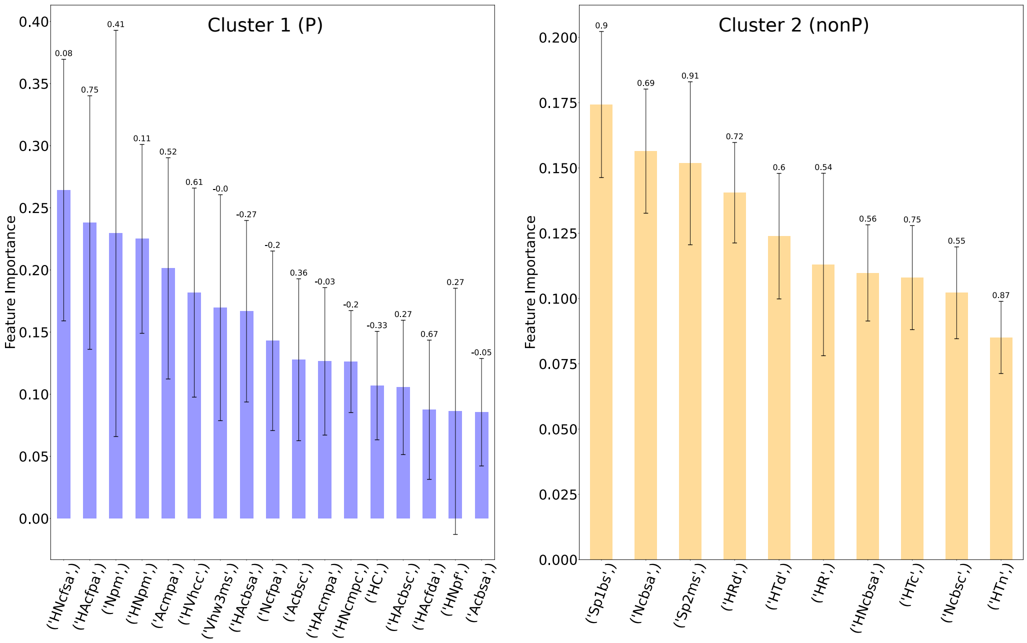


Figure 9: Cross-validated feature importance analysis results for the book of Genesis: **high-res POS** (running-window width of 1; n-gram size of 1). **Note**: displaying features carrying 90% of the explained variance.

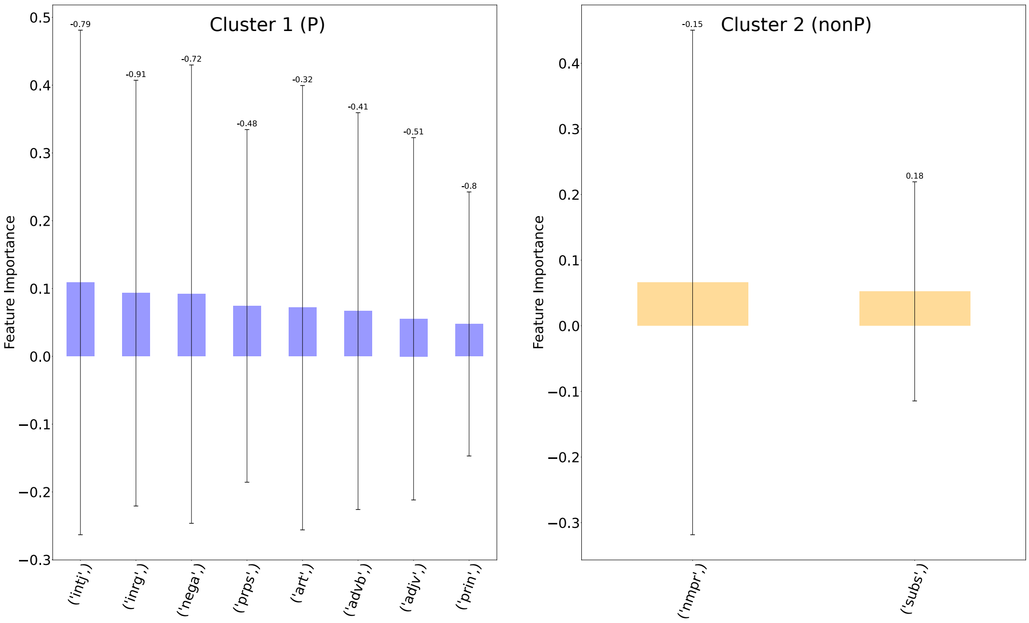


Figure 10: Cross-validated feature importance analysis results for the book of Genesis: **low-res POS** (running-window width of 4; n-gram size of 1). **Note**: displaying features carrying 100% of the explained variance.

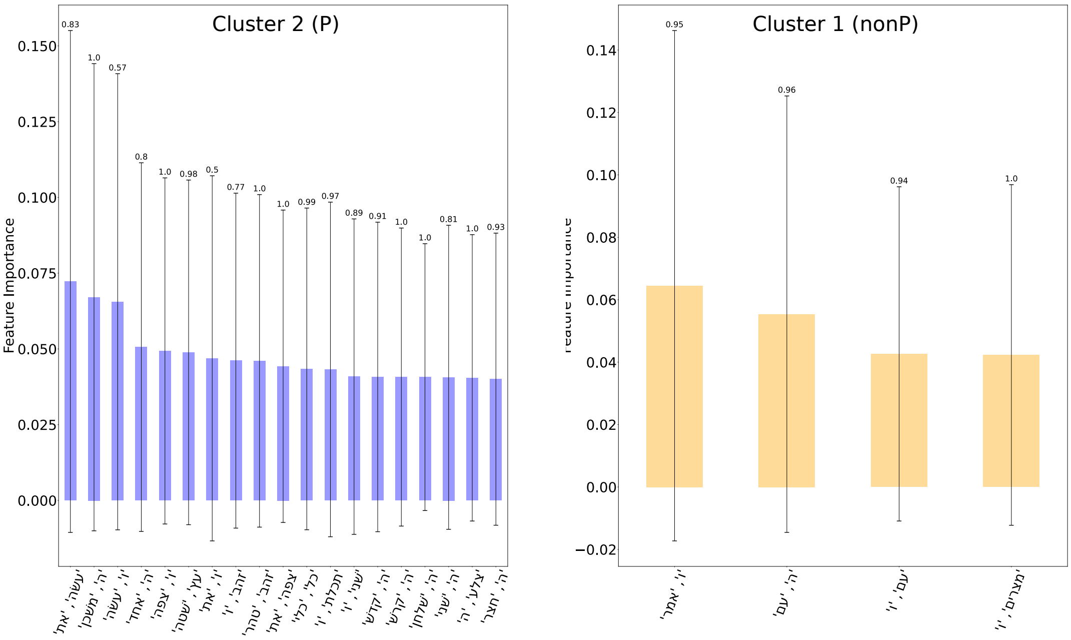


Figure 11: Exodus cross-validated feature importance results: **lexemes** (running-window width of 5; n-gram size of 2).

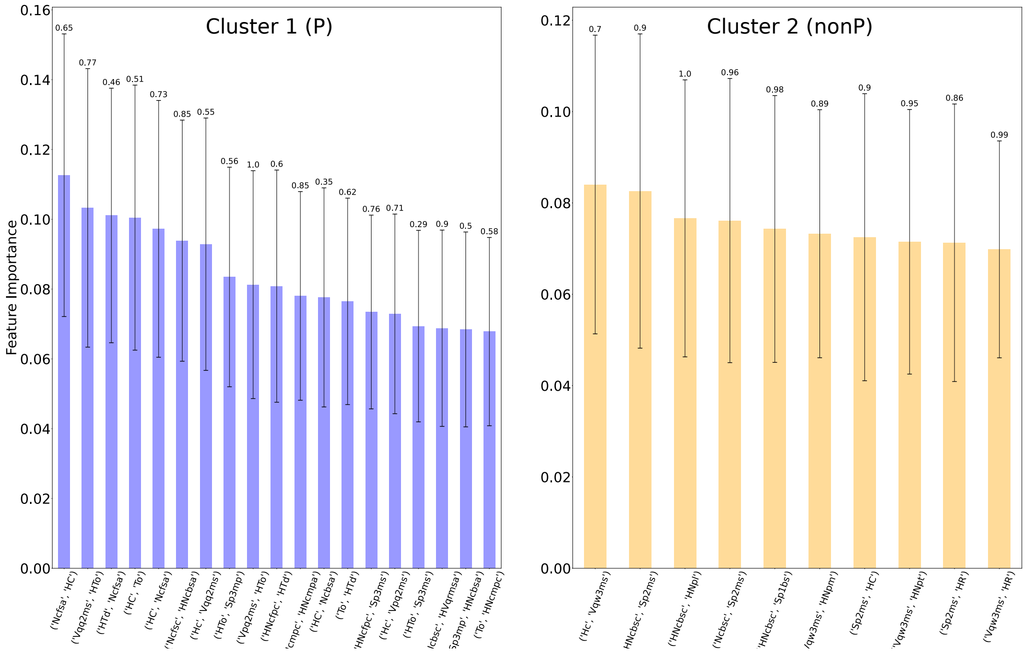


Figure 12: Exodus cross-validated feature importance results: **high-res POS** (running-window width of 4; n-gram size of 2).

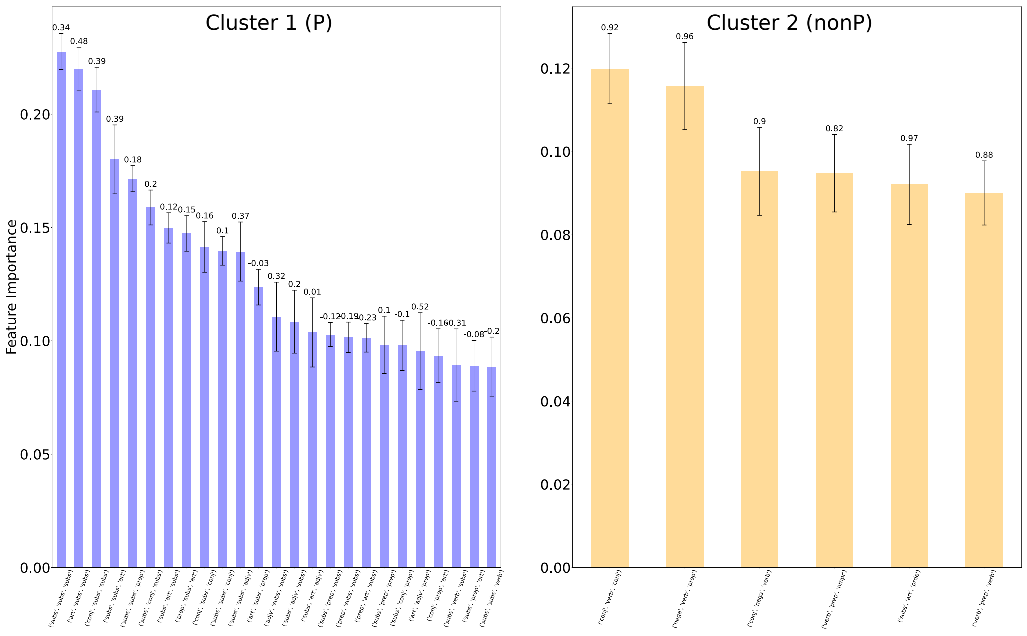


Figure 13: Exodus cross-validated feature importance results: **low-res POS** (running-window width of 6; n-gram size of 3).

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

Left blank.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.