ZfdG

Zeitschrift für digitale Geisteswissenschaften

Fachartikel aus:
Zeitschrift für digitale Geisteswissenschaften, Heft 11 (2026)

Titel:
Applying Text Reuse Detection to Trace Intertextuality in Tafsir Texts

Autor*in:
Thomas Jurczyk

Kontakt: thomas.jurczyk-q88@rub.de
Institution: Ruhr-Universität Bochum (RUB), Centrum für Religionswissenschaftliche Studien (CERES)
GND: 106744744X ORCID: 0000-0002-5943-2305
Contribution (CRediT): Conceptualization | Writing – original draft | Software | Data curation

Autor*in:
Johann Buessow

Kontakt: johann.buessow@rub.de
Institution: Ruhr-Universität Bochum, Institut für Arabistik und Islamwissenschaft
GND: 1014498953 ORCID: 0000-0001-9108-4569
Contribution (CRediT): Writing – original draft | Formal analysis | Project administration

Autor*in:
Roman Seidel

Kontakt: roman.seidel@rub.de
Institution: Ruhr-Universität Bochum (RUB), Institut für Arabistik und Islamwissenschaft
GND: 1073839478 ORCID: 0000-0002-6053-1650
Contribution (CRediT): Writing – original draft | Conceptualization | Formal analysis

DOI des Beitrags:
10.17175/2026_015

Nachweis im OPAC der Herzog August Bibliothek:
1982703415

Erstveröffentlichung:
19.11.2026

Lizenz:

Sofern nicht anders angegeben (cc) BY-SA

Letzte Überprüfung aller Verweise:
01.09.2026

Format:
PDF ohne Paginierung, Lesefassung

GND-Verschlagwortung:
Intertextualität | Metapher | Tafsir | Textvergleich | Islamwissenschaft

Empfohlene Zitierweise:
Thomas Jurczyk / Johann Buessow / Roman Seidel: Applying Text Reuse Detection to Trace Intertextuality in Tafsir Texts. In: Zeitschrift für digitale Geisteswissenschaften 11 (2026). 19.11.2026. HTML / XML / PDF. DOI: 10.17175/2026_015

Thomas Jurczyk / Johann Buessow / Roman Seidel

Applying Text Reuse Detection to Trace Intertextuality in Tafsir Texts

Abstract

This article illustrates how *text reuse detection* can reveal intertextuality and metaphor dynamics in Islamic Qur'anic exegesis (Tafsir), building on a previous *text-mining* study conducted as part of CRC 1475 ›Metaphors of Religion‹. The case study centers on an extensive cluster generated by the *Passim* algorithm documenting the long-term reuse of shared exegetical material from the 11th to the 20th century. The analysis shows how exegetes expanded a specific Qur'anic scene in Q 37:23 through repeated adoption, condensation, and reconfiguration into an elaborate eschatological narrative.

Aufbauend auf einer früheren *Text-Mining*-Studie im Rahmen des SFB 1475 ›Metaphern der Religion‹ zeigt der Beitrag, wie *Text Reuse Detection* bei der Erforschung von Intertextualität und Metapherndynamik in der islamischen Koranexegese (Tafsir) als heuristisches Tool genutzt werden kann. Die Fallstudie untersucht ein umfangreiches, von *Passim*-Algorithmus generiertes Cluster, das die langfristige Wiederverwendung gemeinsamen exegetischen Materials vom 11. bis zum 20. Jahrhundert dokumentiert. Die Analyse zeigt, wie Exegeten eine bestimmte koranische Szene in Sure 37:23 durch wiederholte Übernahme, Kondensierung und Neukonfiguration zu einer ausgearbeiteten eschatologischen Erzählung erweiterten.

1. Introduction

This article extends research on the Tafsir (Qur'anic exegesis) corpus initiated under the CRC 1475 ›Metaphors of Religion‹ project.[1] One of the major aims of the CRC 1475 is to compile a corpus of religious texts from different historical and geographical periods and subject them to metaphor-related research.[2] In the first exploratory study of our subproject A01 within the CRC 1475,[3] we showed how digital methods and tools can support metaphor-related research by applying *text mining* to the Tafsir data. This analysis uncovered semantic clusters in the Tafsir commentaries on Qur'anic verses that use WAY metaphors, in particular those concerning »guidance on the straight and solid road«[4] (***al-ṣirāṭ al-mustaqīm***).[5] [1]

This article analyzes the dynamics of an unconventional metaphor in our Tafsir corpus. Building on previous work, we use computer-driven *text reuse detection* to conduct metaphor-related searches in the Tafsir data. Specifically, we apply text reuse detection to 64 subchapters in our database that discuss Q 37:23,[6] a verse identified in our earlier text-mining study as central to our research on WAY metaphors.[7] Unlike our previous text-mining research, which identified similar topics across Tafsir subchapters without regard to historical context, the present article uses text reuse detection to trace direct relations between authors and texts. This method is more fine-grained and methodologically constrained, as it focuses on the repetition of phrases rather than semantic similarities in larger text units, enabling us to uncover the historical formation, discussion, and evolution of WAY metaphors through intertextual relations. [2]

---

[1] Subproject A01 ›Religious Meaning-Making in the Medium of Space: Close and Distant Read-ings of the Qur'anic Metaphorical Landscape in Tafsir‹ of the Collaborative Research Center 1475, ›Metaphors of Religion‹ (last accessed: May 22, 2026).

[2] Cf. Gebhard et al. 2023.

[3] Jurczyk et al. 2025.

[4] For the early history of the Qur'anic metaphor and a discussion of the translation of *ṣirāṭ* as »road«, cf. Seidel / Buessow 2025a.

[5] Arabic quotations, Arabic book titles, and author names in bibliographical entries were transliterated according to the conventions of the *International Journal of Middle East Studies (IJMES)*. In the running text, author names appear in simplified transcription without macrons or underdots, while retaining the *ʿayn* and the *hamza*. Common Arabic terms, such as Qurʾan, Tafsir, and Hadith, are likewise given in simplified form. Other Arabic terms, including *ṣirāṭ*, *ṣirāṭ al-jaḥīm*, *hidāya*, and *walāya*, are transliterated according to full *IJMES* conventions.

[6] مِن دُونِ ٱللَّهِ فَٱهْدُوهُمْ إِلَىٰ صِرَٰطِ ٱلْجَحِيمِ »instead of God, then lead them all to the road of Hell.«

[7] Cf. Jurczyk et al. 2025.

Although the term *intertextuality* has a complex and extensive intellectual history,[8] we define it here as any explicit reference by one text to another. Text reuse detection allows us to search automatically for various types of intertextuality, ranging from direct quotes and near-verbatim text reproduction to looser forms of reuse, including condensation, selective reuse, and some paraphrastic reformulations. For text reuse detection, we used *Passim*[9] to analyze the Arabic texts. Although newer tools like *alNaql*[10] exist, *Passim* is particularly valuable when the goal is to compare multiple texts and create clusters to chart chronological text reuse in various texts and over time. [3]

Beyond presenting results related to our metaphor-focused research, this article also demonstrates how existing text reuse tools and preprocessing techniques can be applied to Arabic texts. Our approach builds extensively on previous work in this field, especially the *KITAB* project.[11] More broadly, the application of the *Passim* algorithm to 64 Tafsir texts covering the time period from circa 700 to 2000 CE, aims at contributing new insights to Islamic intellectual history. For religious historians, this algorithmic text reuse analysis complements traditional close reading. It maps 13 centuries of exegetical knowledge transmission and thus may help pinpoint hitherto unidentified networks of citations across generations of Qur'anic commentary. Processing this massive Arabic corpus also innovates by quantifying long-term intertextuality at scale, by bridging algorithmic distant reading with a close reading of theological arguments to better chart the evolution of Islamic thought. [4]

2. Brief Introduction of Central Terms

Before presenting this case study, we provide a brief introduction to the Tafsir corpus, the notion of intertexuality, and the methods used here to detect text reuse. [5]

2.1 The Tafsir Corpus

Tafsir is the term for the classical genre of Qur'anic exegesis in which linguistic analysis, transmitted prophetic and early Muslim traditions and rational argumentation are used to explain the meanings, contexts, and implications of the Qur'an. There is extensive literature on Tafsir as a genre,[12] its history and its doctrinal diversification,[13] and the current state of the art in research.[14] For the purposes of this study, the metatextual structure[15] underpinning all examples of the Tafsir genre in its traditional sense is of prime importance. This structure, which is characterized by verse-by-verse commentary of the Qur'anic texts, is particularly well-suited to a comparative and diachronic analysis of a large corpus of texts (the Tafsir works) that shares a single earlier work to which they refer (the Qur'an). [6]

[8] Cf. Allen 2019.
[9] Cf. Romanello / Hengchen 2021.
[10] Cf. Kelly 2025. Although there are other available text reuse detection tools such as TextPAIR, *BLAST* (cf. Vesanto et al. 2017), *Tesserae* (cf. Coffee et al. 2013), and *Tracer* (cf. Franzini et al. 2018), scholars working on Arabic texts tend to use *Passim* and the recently published *alNaql*. *AlNaql* has useful Arabic-specific features, such as *isnad* filters and various similarity metrics. *Isnad* is the (oftentimes very long) chain of transmitters (narrators) through which a *hadith* (a saying or action of the Prophet Muhammad) is reported. However, *alNaql* currently cannot process more than two texts simultaneously or create text reuse clusters.
[11] ›KITAB – Knowledge, Information Technology, & the Arabic Book‹. The project, directed by Sarah Bowen Savant, is funded by the European Research Council (KITAB, no. 772989) and the Qatar National Library (KITAB, last accessed: March 10, 2026).
[12] Görke / Pink (eds.) 2014; Pink 2020; Pink 2010; Sinai 2017.
[13] Gilliot 2002; Rippin 2000; Shah / Abdel Haleem (eds.) 2020.
[14] Ross 2023.
[15] Cf. Genette 1997.

Our Tafsir corpus is comprised of 80 Tafsir works originally taken from the website Altafsir.com. The texts were first scraped by Dr. Alicia Martínez González, refined, and compiled into a JSON corpus, which was then integrated into the larger *OpenITI* corpus.[16] Next, this JSON corpus was processed and enriched with metadata and linguistic information by Adrian Bernhard in his master's thesis.[17] The corpus is continually updated and expanded as part of the ongoing A01 project.[18] [7]

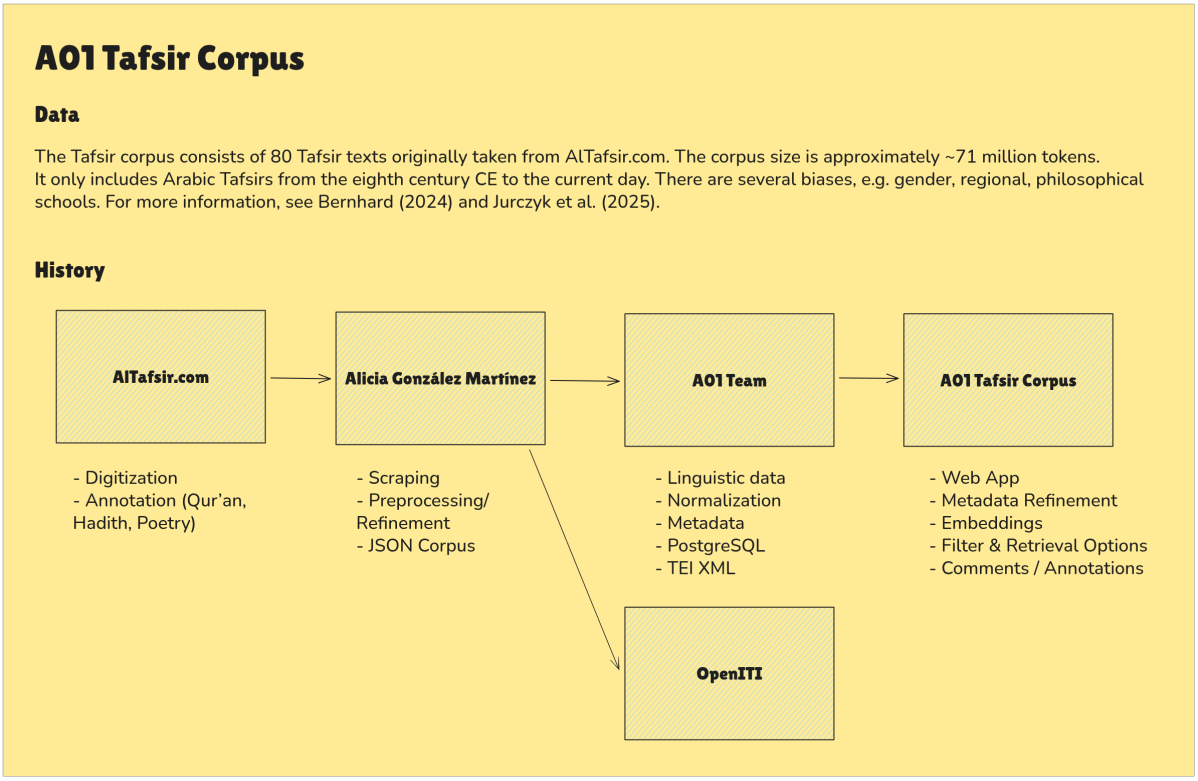


Figure 1: Genesis of the Tafsir corpus. [Diagram: Thomas Jurczyk 2026]

The Tafsir corpus has a total of 71,613,695 tokens.[19] It includes works that vary enormously in length and range from over 3 million tokens (al-Ālūsī, *Rūḥ al-Maʿānī*) to about 169,000 tokens (ʿAbd al-Salām, *Tafsīr*), some of which are incomplete. The corpus contains Arabic Tafsir texts dating from approximately the eighth century CE (based on the death dates of the authors attributed to the works) up to recent times (cf. Fig. 2). In terms of regional distribution, it primarily consists of texts from the Middle East. Major regions such as India and Southeast Asia are underrepresented. [8]

[16] Nigst et al. 2023.
[17] Bernhard 2024.
[18] The corpus is currently accessible via the Tafsir app, a web-based research interface developed as part of the project (last accessed: February 22, 2026).
[19] For a detailed overview of the corpus, cf. CRC-1475 (last accessed: March 10, 2026).

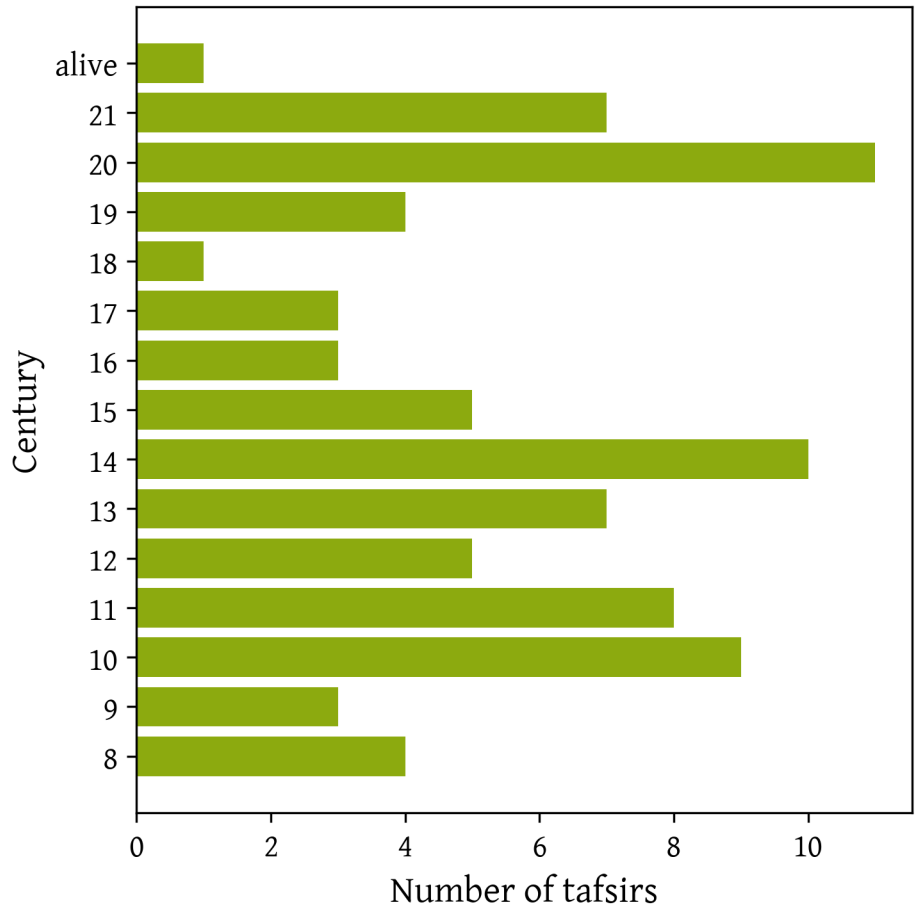


Figure 2: Tafsirs by century (according to their authors' assumed death dates). [Source: Bernhard 2024, p. 81]

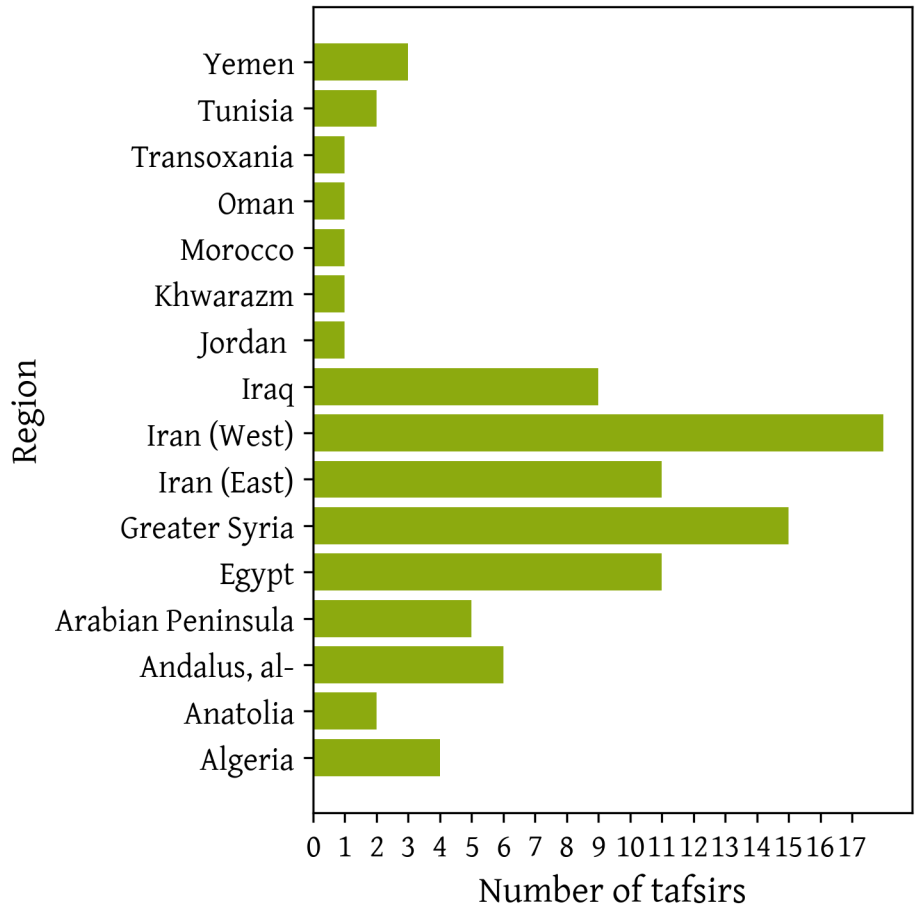


Figure 3: Tafsirs by region. [Source: Bernhard 2024, p. 82]

These regional and chronological biases are inherent to the original Altafsir.com collection, despite its unusually broad scope and stated aim to be ecumenical, inclusive, comprehensive, and ideologically independent encompassing »standard Classical and Modern Commentaries [...] of all eight schools of jurisprudence.«[20] These structural biases can largely be explained by the vast number of Tafsir texts that remain unedited and undigitized.[21] Nevertheless, despite these limitations, to the best of our knowledge, the current corpus is the largest digitized, structured, and (partially) annotated collection of Tafsir texts available to date and thus provides an excellent foundation for computational analysis. [9]

In addition to regional and chronological information, the curated Tafsir corpus prepared on the basis of Altafsir.com includes metadata categories selected for their analytical relevance to exegetical transmission: theological affiliation, legal school, and dominant exegetical concerns. These categories help contextualize how commentators adopt, modify, or reject inherited material and allow us to relate patterns of text reuse to broader intellectual, doctrinal, and methodological orientations.[22] [10]

[20] Bernhard 2024, p. 48.
[21] Cf. Ross 2023.
[22] The assignment of metadata categories to individual Tafsir works is based primarily on what is known about their authors rather than, in every case, on securely datable features of the texts themselves. For example, chronological classification relies on the author's death date, since the date of completion of a given commentary is often unknown.

In addition to these categories, the Tafsir corpus includes annotations of Qur'an, Hadith,[23] and poetry quotations in the Tafsir texts. These annotations were manually created by the Altafsir.com team and provide an invaluable resource for advanced filtering, text retrieval (e.g., searches with or without quotations), and statistical analyses (for instance, identifying which Tafsirs contain the highest number of poetry quotations[24]). [11]

In terms of the linguistic data, the entire corpus was processed using CAMeL Tools.[25] Although primarily designed for Modern Standard Arabic (MSA), it enabled us to extract valuable linguistic information from the historical Arabic texts at the word level, including (de-)diacritized and stemmed forms. These outputs were used to generate customized versions of the Tafsir data for subsequent processing, such as producing a lemmatized version of the texts for semantic analysis. [12]

The corpus is currently maintained and updated by the A01 project, which is adding additional functionalities such as text embeddings, automated translation features, and advanced filtering and retrieval options through its internal corpus tool (cf. Fig. 4). Both the final dataset and the web application are planned for public release at a later stage. [13]

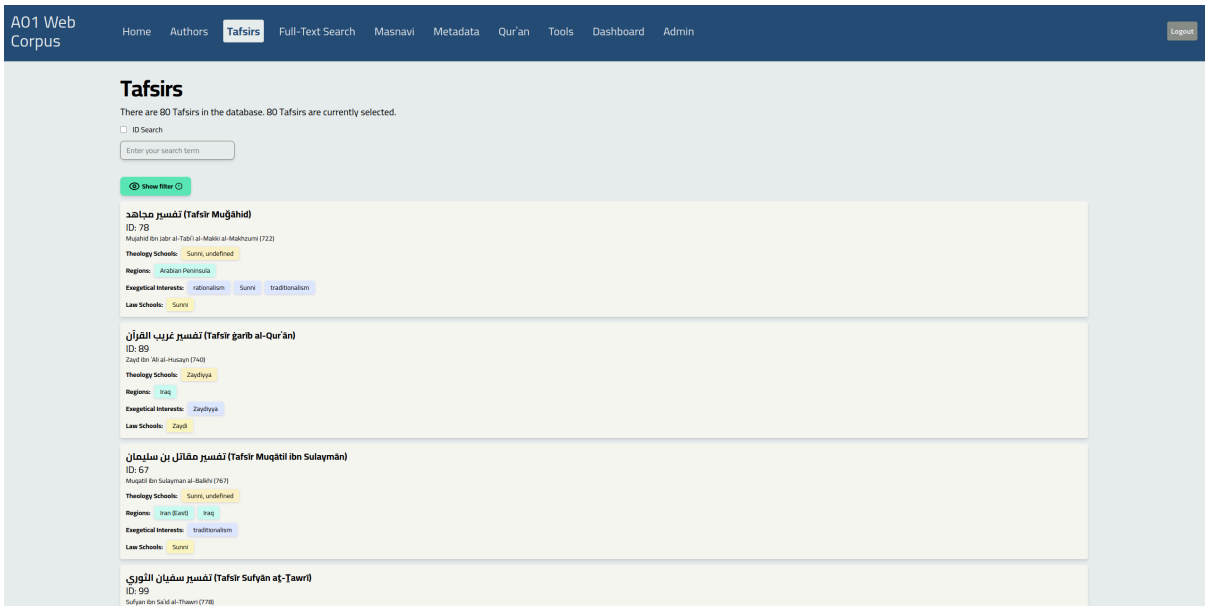


Figure 4: Screenshot of the A01 Tafsir application, showing an overview of Tafsir texts with metadata annotations. [Source: Tafsir app]

2.2 Intertextuality

To conceptualize intertextuality as part of the broader notion of transtextuality, we drew on the work of the literary theorist Gérard Genette[26]. Genette coined the term ›transtextuality‹ in his work *Palimpsests: Literature in the Second Degree*, where it is defined as any relationship between a text and one or more other texts. This concept is broken down into five categories, each with its own subdivisions: [14]

1. ›Intertextuality‹, which consists of the direct presence of one text within another, such as quotes or plagiarism. In the context of Tafsir, this includes any type of quotation.

[23] A *hadith* is a recorded saying, action, or approval of Muhammad, serving as a primary source of guidance in Islam alongside the Qur'an.
[24] See also Jurczyk et al. 2025 for preliminary studies in this direction.
[25] Obeid et al. 2020.
[26] Cf. Genette 1997.

2. ›Paratextuality‹, which refers to any text that frames another text, such as a preface, titles, marginal notes, or afterwords.
3. ›Metatextuality‹, which consists of instances where one text comments on another. As Qur'anic commentary, Tafsir is a clear example of metatextuality.
4. ›Architextuality‹, which describes the relationship of a text to other texts of its kind, such as the connection between a work and its genre. In our context, this involves the relationship between a Tafsir work and the Tafsir genre.
5. ›Hypertextuality‹, which encompasses any relationship that links a later text (hypertext) to an earlier text (hypotext). Genette primarily concentrated on texts that transform or rewrite previous texts, with the exception of commentaries. In the context of the Tafsir texts, the Qur'an can be considered the hypotext, whereas the Tafsir is the metatext or, more broadly, the hypertext.

[15] This study deals primarily with identifying instances of text reuse, especially quotations and paraphrases, which correspond to intertextuality. At the same time, these intertextual relations are embedded in a broader transtextual constellation. Tafsir is metatextual insofar as it comments on the Qur'an, and hypertextual insofar as it transforms, expands, and reconfigures Qur'anic meanings. The analysis of text reuse in different Tafsir works can therefore not only provide insights into intertextual links, but also into more pervasive processes of metatextual interpretation and hypertextual transformation within the exegetical tradition.[27]

2.3 Metaphors

[16] In the Islamic religious tradition, metaphors of space and movement are particularly important, beginning with the »straight and solid road« (*al-ṣirāṭ al-mustaqīm*) mentioned in the opening chapter of the Qur'an. Because the overall research project investigates how metaphorical language in Tafsir contributes to the production of religious meaning and articulates different religious imaginaries,[28] metaphors are not treated as rhetorical embellishments, but as central devices that enable religious concepts to become cognitively accessible, textually structured, and historically transmitted.

[17] Our approach is grounded in the distinction between ›Qur'anic lead metaphors‹ and ›explicative metaphors‹. Qur'anic lead metaphors are recurring conceptual metaphors that structure the Qur'an's literary and religious world. These are expressed through multiple textual metaphors, and function as focal nodes within networks of interrelated images. Explicative metaphors are exegetical metaphors found in Tafsir and other hypertexts related to the Qur'an. They elaborate on Qur'anic lead metaphors by unfolding their meaning, enriching their imagery, and sometimes shifting their interpretive focus. In this sense, Tafsir does not merely explain Qur'anic metaphors; it can also concretize, disambiguate, redirect, or transform them. Our analytical process thus took place in two stages: the first involved identifying and analyzing Qur'anic lead metaphors, and the second consisted of tracing their explicative development historically in the Tafsir corpus.

[18] This process also clarified the value of text reuse detection for metaphor analysis. Text reuse can reveal how exegetical formulations percolate across works, periods, regions, exegetical approaches, and theological schools of thought. When combined with metadata such as chronology, provenance, school affiliation, and / or exegetical orientation, text reuse detection can be used to reconstruct not only textual dependencies but also the historical trajectories of metaphorical interpretation. It thus functions as a heuristic tool for identifying where metaphorical innovations occur, where and when interpretations become conventionalized, and which alternative readings remain latent or reappear in later traditions.

[27] For a more in-depth theoretical presentation, cf. Seidel 2026.
[28] For more on methodological approaches to annotating and analyzing metaphors in Qur'an and Tafsir, cf. Seidel / Buessow 2025b.

To illustrate this process, we present a case study that centers on an unusual Qur'anic occurrence of the *sirat* metaphor: Q 37:23, where the condemned are commanded to be led to »the road of Hellfire« (*jaḥīmṣirāṭ al-jaḥīm*). This formulation is striking because it inverts the dominant Qur'anic use of *sirat* as a metaphor for guidance, righteousness, and salvation. Rather than evoking the »straight and solid road« leading toward God, the expression implies chastisement. We examined how commentators interpret *al-ṣirāṭ al-mustaqīm* and whether this inverted occurrence has given rise to other explicative metaphors within Islamic exegetical literature. [19]

This case is especially well-suited to a study of text reuse. On one hand, the metaphor is lexically traceable through a clearly identifiable metaphor-related word, *sirat*. On the other, its interpretation is conceptually unstable: exegetes can interpret it as a road, an eschatological bridge, or as an ambiguous image in which both conceptualizations remain available. This combination of lexical stability and semantic openness makes Q 37:23 a productive test case for examining how text reuse participates in metaphor development in Tafsir works. It allows us not only to inquire which passages are reused, but also how the reused material is condensed, expanded, reordered, and / or re-semanticized across changing exegetical contexts. [20]

2.4 Text Reuse Detection

We implemented text reuse detection to automatically identify intertextuality between texts. Text reuse detection involves comparing two or more documents – either in full or by identifying similar passages – through text alignment techniques. Text alignment also serves as a key tool for other humanities tasks such as text collation, synoptic editions, and linking original texts to their translations.[29] In addition to humanities research, text reuse detection and alignment are also important to areas such as plagiarism detection. [21]

According to Miller et al.,[30] text reuse detection typically involves three key steps: [22]

1. Identification of document pairs, which consists of identifying pairs of documents that are likely to contain instances of text reuse. This step is particularly important when dealing with large amounts of textual data where only a few candidates for text reuse are expected and a comparison of all textual pairs would exceed computational limitations. Potential methods to identify relevant document pairs are, among others, the analysis of n-gram overlaps between texts[31] and *TF-IDF (Term Frequency – Inverse Document Frequency)* approaches.[32]
2. Application of text alignment algorithms to preselected documents. The goal of this step is to identify passages in the document pairs that include text reuse. These texts can be divided into smaller segments as a function of their length.
3. Statistical evaluation to assess the extent of text reuse between documents based on the text alignment.

Tariq Yousef and Stefan Jänicke[33] expanded on the Miller et al. approach to detail the steps involved in text alignment: [23]

1. Preprocessing of the (preselected) texts, which, depending on the length of the documents, can consist of splitting the documents into smaller parts, such as paragraphs, sentences, or even words. This step can also include further preprocessing, such as stemming, setting the text to lowercase, or the removal of unwanted characters.

[29] Cf. Brill et al. 2020; Yousef / Jänicke 2021, pp. 1145–1150.
[30] Miller et al. 2025.
[31] Cf. Clough et al. 2002, p. 154.
[32] Cf. Sturgeon 2018.
[33] Yousef / Jänicke 2021, pp. 1152–1153.

2. Text alignment yielding text unit pairs of matching texts based on the selected text alignment algorithms (e.g., Smith–Waterman and Needleman–Wunsch; see Section 2.4.2).
3. Visualization of the results as sequence-based, grid-based, text-oriented heat maps, or aligned barcodes to facilitate interpretation.[34]

Visualizing the results is crucial when using text reuse detection as a heuristic tool or when dealing with a large volume of texts. There are numerous approaches to visualizing text-reuse data, depending on the number of texts being compared (one-to-one, one-to-many, many-to-many).[35] A good example of a one-to-one text-reuse visualization is provided by the interactive plots developed by the *KITAB* project (cf. Fig. 5). These visualizations are based on the 300-word chunks that *KITAB* uses throughout its extensive *OpenITI* library[36] which make it possible to visually trace the magnitude and the distribution of text reuse between two works from different perspectives. On cursory examination, the amount of red directly indicates the overall extent of text reuse relative to the size of the works. For a more detailed analysis, the interactive plot enables users to trace exactly where each reused passage appears in the other text. The plot's interactivity also allows users to open the full texts at the points where reuse occurs. Thus text-reuse visualizations, when properly designed, are more than alternative views of the data: they enable users to examine the data in multiple ways, some of which would be difficult or take much longer without the algorithm presentation. [24]

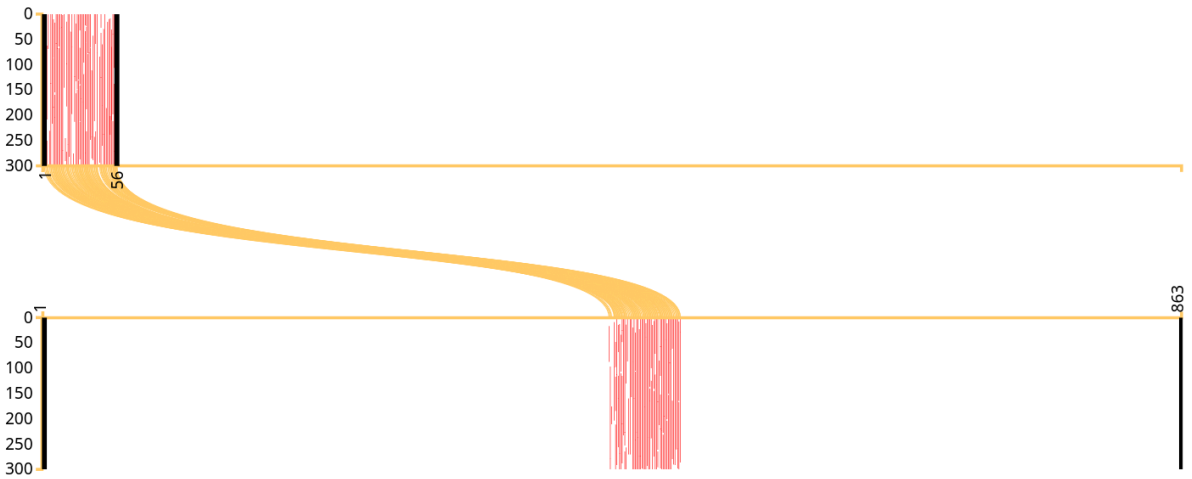


Figure 5: *KITAB* visualization of one-to-one text reuse in ʿAntara Ibn Shaddād, *Diwan* and Luwīs Shīkhū al-Yasūʿī, *Shuʿarāʾ al-naṣrāniyya baʿda l-islām* . [Source: KITAB]

KITAB also lets users choose from other representation techniques for one-to-many comparisons, such as dotplots (cf. Fig. 6).[37] Other popular alternatives for the visualization of more than two texts[38] are text-reuse grids and text-oriented heat maps.[39] [25]

[34] Cf. Buschmann / Fink 2023; Yousef / Jänicke 2021.
[35] Cf. Yousef / Jänicke 2021; Jänicke et al. 2014.
[36] *KITAB* uses *Passim* for text reuse detection, which was initially designed for smaller news corpora. In order to work with the extensive *OpenITI* corpus, *KITAB* splits the texts into 300-word chunks which are then used as input to the *Passim* algorithm. For details, see the KITAB website (last accessed: February 15, 2026).
[37] Cf. Buschmann / Fink 2023.
[38] Cf. Yousef / Jänicke 2021, p. 1151.
[39] Cf. Yousef / Jänicke 2021, p. 1154.

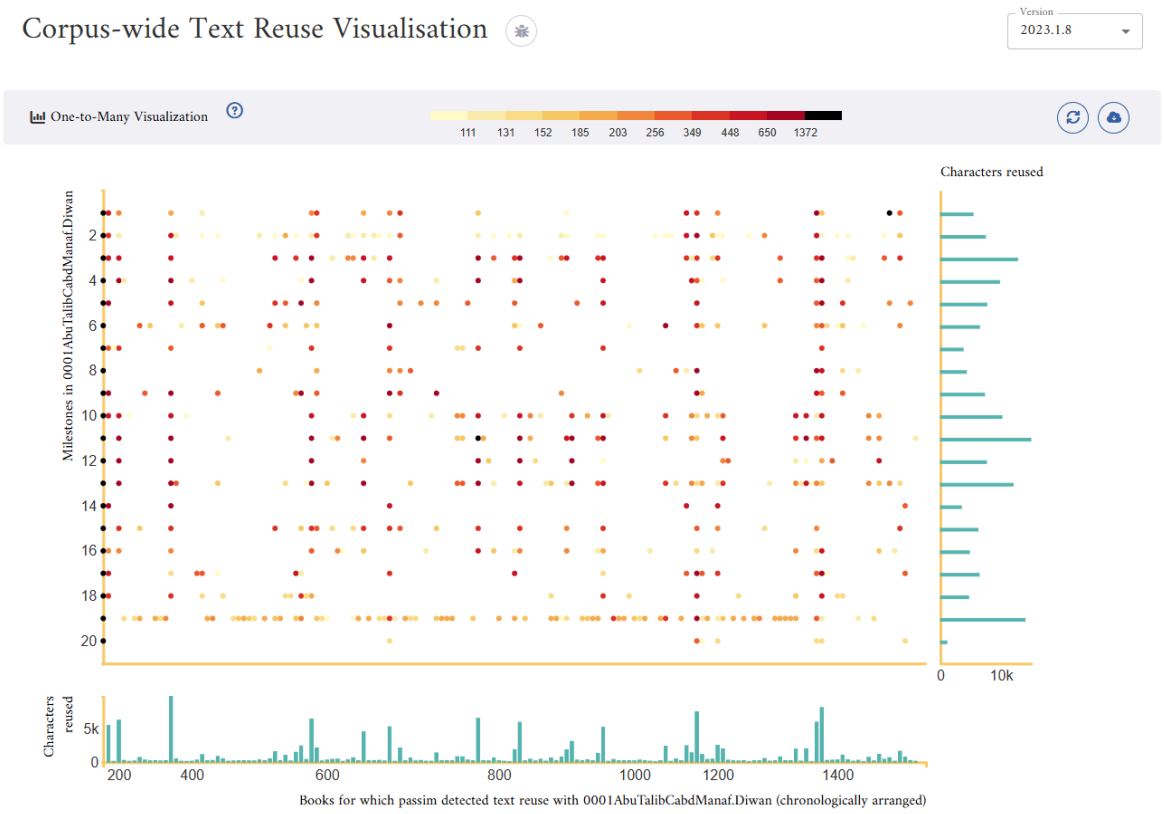


Figure 6: *KITAB* corpus-wide one-to-many visualization of Abu Talib ʿAbd Manaf, *Diwan*. [Source: KITAB]

Since this article focuses on chronologically tracing text reuse across relatively small texts (Tafsir subchapters), which involves a many-to-many relationship, we chose to implement *directed acyclic graph (DAG)* visualizations to represent the data. One example of a graph-based visualization for chronologically ordered documents is the interactive interface developed by the ›Graph – Text Reuse in Rare Books‹ project (cf. Fig. 7). A related, though non-chronological, graph-based approach was developed in the ›Viral Texts‹ project.[40] [26]

[40] Cordell / Smith 2024, Network of »Viral Text« Sharing, 1836–1860 (last accessed: August 18, 2025).

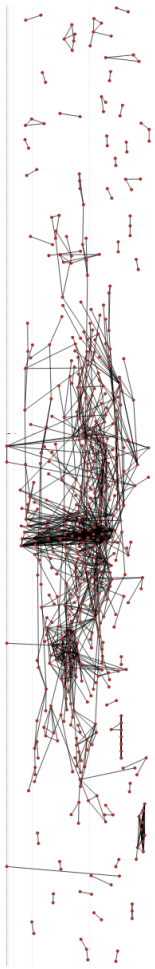


Figure 7: Chronological text reuse in the form of an acyclic directed graph (DAG). [Source: Graph – Text Reuse in Rare Books]

For our project, we developed a similar graph-based visualization on our project website (cf. Fig. 8),[41] which enables users to generate graph views alongside the *Passim* clusters identified in the Tafsir data, which can be particularly helpful in larger clusters to get a better understanding of the chronological distribution. [27]

[41] A01 Text Reuse Application (last accessed: March 10, 2026).

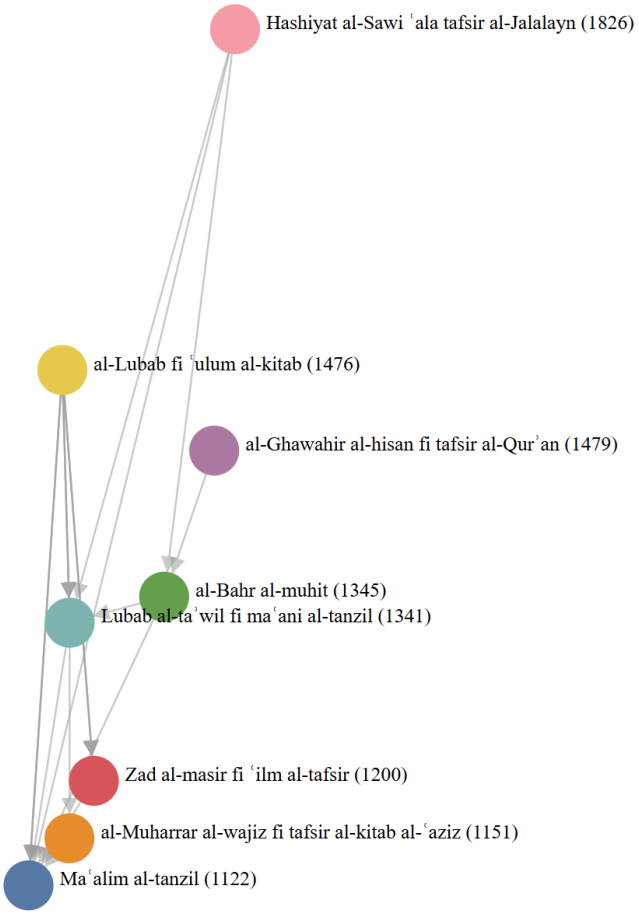


Figure 8: Exemplary DAG visualization of a *Passim* cluster on our internal text-reuse detection website. [Interactive graph: A01 Text Reuse Application, generated with D3.js]

2.4.1 Passim

One widely-used implementation of the three-fold text reuse detection approach is *Passim*[42] which is increasingly being used as a detection solution in many humanities-related projects, such as *KITAB*[43] and *impresso*[44], and is actively maintained[45]. *Passim* starts by identifying suitable text-reuse candidate pairs, a critical step that reduces the number of pairs processed during the computationally expensive text alignment phase. This process involves the following sequence: [28]

1. Define the n-gram length to identify n-gram matches between documents. In recent versions of *Passim*, this n-gram is based on characters, not tokens.
2. For each n-gram that appears in more than one document, create a list of (document ID, n-gram position) tuples.
3. Extract all unique document pairs of each n-gram from that list.
4. Set a value (m) to be the minimum threshold of unique n-gram matches between two documents. If documents share the minimum number of n-gram matches, they are regarded as potential text-reuse candidates.

[42] Cf. Romanello / Hengchen 2021; Smith et al. 2015.
[43] Cf. Barber 2020.
[44] Cf. Düring et al. 2023.
[45] This article is based on *Passim* version 2.0.1, which was released in February 2024 (Smith 2015).

Then, a text-alignment algorithm is applied to the list of text-reuse candidate segments containing the matching n-grams. Neighboring text-reuse passages are merged based on a gap parameter. In earlier iterations of *Passim*, this step required the Smith–Waterman alignment algorithm. Versions 2.0.1 and above include a more elaborate and complex dynamic-programming local alignment algorithm. [29]

Clusters are formed by treating each aligned passage as a node in a graph, with edges connecting the alignment pairs across documents. Since the clusters are derived from the transitive closure of these edges, documents can be grouped together even without direct pairwise alignment. For example, if a passage in document A aligns with B, and the same (or overlapping) passage in B aligns with C, then A, B, and C share a cluster, even if A and C were never directly compared. This property is especially useful for tracing chains of text reuse over time, where a passage can spread from an early source through multiple intermediate texts. [30]

We applied *Passim* to a version of the Tafsir subchapters that had undergone several preprocessing steps including chunking and the removal of Qur'anic quotes which were trivially responsible for most text-reuse cases in subchapters that all comment on the same Qur'anic verses. Their removal made it possible to identify instances of genuine text-reuse rather than overlaps caused simply by shared scriptural references. [31]

2.4.2 Text Alignment Approaches

Text alignment algorithms aim to identify similar documents or passages within different documents through various methods of automated text comparison. Many of these algorithms operate using matrices, where the elements being compared (such as characters, tokens, or sentences) from two documents, A and B, are listed as indices in the rows and columns of the matrix. The cells within the matrix then indicate the alignment or similarity between the elements of documents A and B. [32]

For example, for document A with the text »The quick brown fox« and document B with the text »The quick red fox«, the value 0 at position (1, 2) in the matrix indicates that the token at index 1 of Document A (»The«) and the token at index 2 of Document B (»quick«) do not align, whereas they align at position (2,2) (»quick« and »quick«). [33]

	The	**quick**	**red**	**fox**
The	1	0	0	0
quick	0	1	0	0
brown	0	0	0	0
fox	0	0	0	1

Table 1: Illustrative matrix-based text alignment table.

Other approaches like *CollateX* utilize different data structures to represent text alignments, such as variant graphs,[46] as shown in Fig. 9 where the empty nodes at the beginning and end indicate the start and end of the variant graph. [34]

[46] Cf. Haentjens Dekker et al. 2015, pp. 459–461.

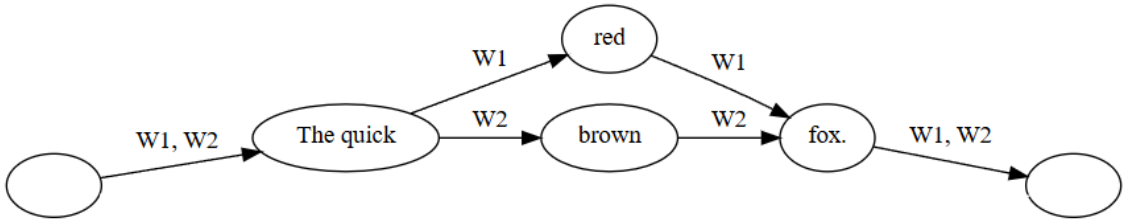


Figure 9: Variant graph for the two documents W1 (»The quick red fox.«) and W2 (»The quick brown fox.«) created using the *CollateX* web interface. [Source: CollateX]

The two most popular matrix-based text alignment algorithms are the Needleman–Wunsch[47] algorithm and the Smith–Waterman[48] algorithm.[49] Needleman–Wunsch emphasizes the similarities between entire texts, whereas Smith–Waterman aims to identify similar passages locally within documents. Many alignment algorithms originate from the field of bioinformatics where they were developed to identify similar DNA or RNA sequences.[50] This helps account for their binarity and rigidity that are not always optimal for texts and humanities data and has consequently led to adoptions and refinements of these approaches for humanities research.[51] [35]

The comparison of individual words or characters can be tricky when the texts contain noisy (OCR) data, misspellings, or differing historical writing conventions. One way to address these variations is to lemmatize or stem the texts beforehand, which can be difficult in the context of some historical languages,[52] or to employ string comparison metrics like the *Levenshtein distance*,[53] a specific type of edit distance that identifies similar character sequences within a word rather than similar word sequences in a document. [36]

However, the Levenshtein distance cannot resolve the problem of different words that have a similar meaning. For example, words such as ›begin‹ and ›commence‹ have a very high Levenshtein distance of seven, but are very similar in meaning. To deal with this issue, attempts have been made to enhance the Needleman–Wunsch and Smith–Waterman algorithms by changing the distance functions between words and the underlying cost functions by including word embeddings as a conceptual layer into the distance calculation between words, sometimes combined with more orthographic string comparison measures such as the Levenshtein distance.[54] [37]

Miller et al. developed a sentence alignment algorithm called *FAST* designed to create synoptic passages between texts.[55] This method involves techniques for dividing larger documents into comparable sections and also combines orthographic (Levenshtein) and conceptual (*Word2Vec*) distances to measure word similarities within a Needleman–Wunsch text alignment framework. Miller et al. suggested a modified version of the Smith–Waterman algorithm (rather than Needleman–Wunsch) that also utilizes a combined distance function.[56] Unlike the Brill et al. approach,[57] this function incorporates *fastText* embeddings as a conceptual layer and the normalized Levenshtein distance for orthographic comparisons. In both [38]

[47] Needleman / Wunsch 1970.
[48] Smith / Waterman 1981.
[49] However, there are other approaches as well, such as *Greedy String-Tiling* (see Clough et al. 2002, p. 155).
[50] Cf. Buschmann / Fink 2023; Yousef / Jänicke 2021, p. 1149.
[51] For more details, cf. Buschmann / Fink 2023.
[52] Cf. Miller et al. 2025.
[53] Levenshtein 1966.
[54] Cf. Haentjens Dekker et al. 2015.
[55] Miller et al. 2025.
[56] Miller et al. 2025.
[57] Brill et al. 2020.

approaches, the integration of embeddings improves the influence of semantic similarity comparisons, rather than relying solely on identifying orthographically similar words, which can be both an advantage as well as a disadvantage depending on the issue at hand.

Overall, both approaches shift the focus in text reuse detection from direct quotations or text reproduction to the identification of paraphrases and allusions, making text alignment and text reuse detection more applicable in contexts where tracing concepts and ideas is more critical than identifying exact text reproduction or collating different versions of the same text. [39]

3. Data

We examined the subchapters discussing Q 37:23 that emerged as relevant for examining WAY metaphors in our earlier text-mining study.[58] To do so, we constructed a version of the subchapter corpus on Q 37:23 including Qur'anic quotations (Corpus QQ) and one without (Corpus NQQ) because Qur'anic quotations were frequently the primary source of *Passim*'s n-gram matches and alignments. Since these quotations, which often vary in length, appear in nearly all subchapters, we decided to create a separate corpus rather than rely on *Passim*'s maximum document-frequency parameter. [40]

The statistical analysis deals exclusively with the subchapter corpus without Qur'anic quotes (NQQ), which provided the most informative results. However, for transparency, we include the statistics for both corpora to show how much text was lost during preprocessing. [41]

Not all Tafsir subchapters address a single verse; in fact, some cover multiple thematically connected verses from the same sura.[59] Combined with the fact that some Tafsirs are only fragmentarily transmitted, this results in considerable variation in the length of individual subchapters, as shown in Fig. 10 and Table 2, ranging from 39 characters to as many as 24,553 characters (excluding whitespaces).[60] [42]

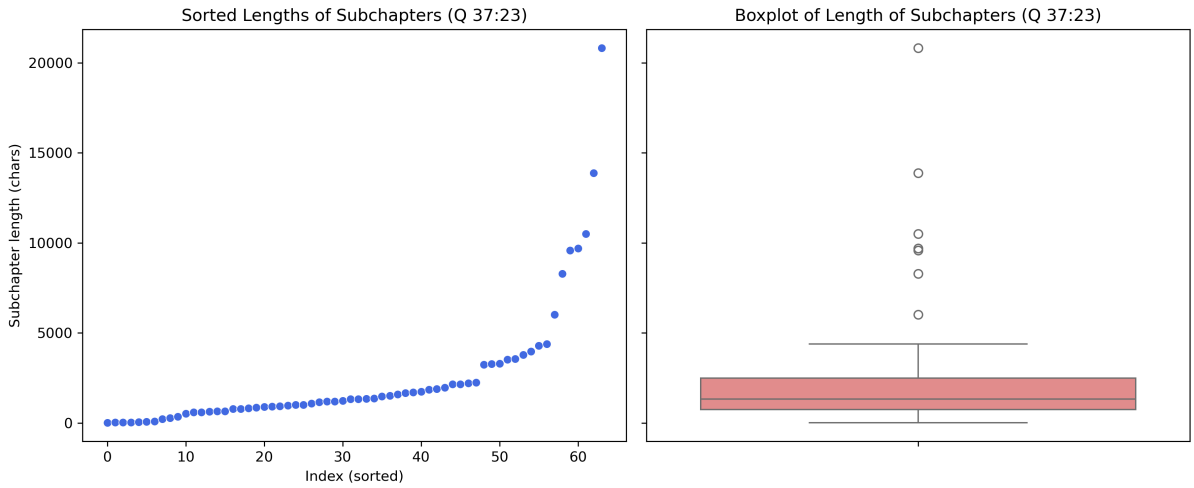


Figure 10: Scatter plot and boxplot of the text lengths (characters) of the preprocessed Corpus NQQ. This corpus was stripped of Qu'ranic quotes and the preprocessing pipeline was applied (including removal of punctuation marks and non-Arabic characters). Whitespaces were also ignored. [Plot: Thomas Jurczyk 2026]

[58] Jurczyk et al 2025.
[59] For example, the Tafsir *Majmaʿ al-bayān fī tafsīr al-Qurʾān* (ca. 767 CE) discusses Q 37:21–30 in the same subchapter (cf. Tafsir app, last accessed: February 22, 2026).
[60] The following exploratory analysis of the subchapters can be found in the Jupyter Notebook *create_chunks_from_subchapters.ipynb* on Codeberg.

	Documents (number)	Avg. chars (mean)	Avg. chars (median)	Max. chars	Min. chars
Corpus (full texts)	64	3,401.7	1,981.5	24,553	39
Corpus (NQQ)	64	2,523.3	1,333	20,818	20

Table 2: Descriptive statistics for the two subchapter corpora. The corpus without Qur'anic quotes was stripped of Qur'anic quotes and the preprocessing pipeline was applied (including removal of punctuation marks and non-Arabic characters). Whitespaces were also ignored.

Since *Passim* was initially designed to handle relatively short texts of similar size such as newspaper articles,[61] we decided to split all the files into 300-word chunks as done in the *KITAB* project, which split its *OpenITI* corpus into 300-word chunks after determining that this was the optimal size for processing Arabic texts with *Passim*.[62] [43]

However, this approach fails to document text reuse that occurs between chunks. Unlike *KITAB*, which divided its corpus into 300-word files and modified the *Passim* algorithm to trace text reuse across these boundaries, we used a floating window approach to create intermediary chunked text files. This method accounts for potential text reuse between the original chunks. Although it increases the overall corpus size and introduces some redundancy in matches (due to overlapping text parts), the impact was negligible given the small corpus of only 64 subchapters. We also removed 5 chunks with fewer than 50 characters (including whitespaces); i.e., below our minimum character length threshold (50 by default).[63] [44]

As shown in Fig. 11 and Table 3, the new chunked data were much more evenly distributed, and provided a better foundation for text reuse detection with *Passim* even prior to excluding texts with fewer than 50 characters. [45]

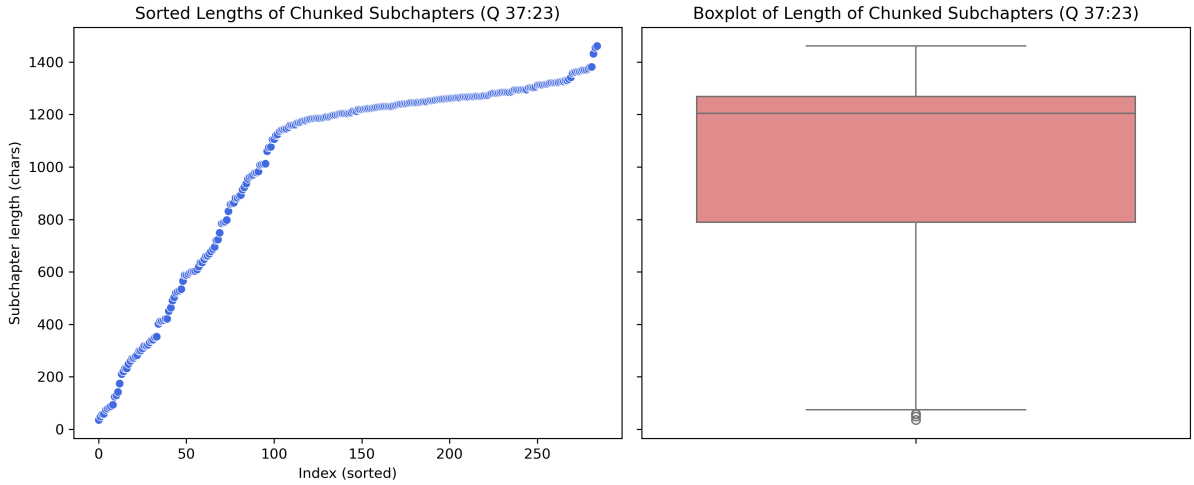


Figure 11: Scatter plot and boxplot of the text lengths (characters) of the preprocessed and chunked Corpus NQQ. This corpus was stripped of Qu'ranic quotes and the preprocessing pipeline was applied (including removal of punctuation marks and non-Arabic characters). Whitespaces were also ignored. [Plot: Thomas Jurczyk 2026]

[61] Cf. Smith et al. 2015.
[62] Cf. KITAB (last accessed: March 2, 2026).
[63] Note that the removal of chunks with less than 50 characters was only done right before the analysis. The data is still visible in the visualizations and the statistical overviews.

	Documents (number)	Avg. chars (mean)	Avg. chars (median)	Max. chars	Min. chars
Chunked Corpus (NQQ)	285	1,011.59	1,205	1,461	36

Table 3: Descriptive statistics for the two subchapter corpora with chunking applied. The corpus without Qur'anic quotes was stripped of Qur'anic quotes and the preprocessing pipeline was applied (including removal of punctuation marks and non-Arabic characters). Whitespaces were also ignored.

3.1 Preprocessing and Workflow

As shown in Fig. 12, the processing pipeline consisted of preparing the texts, running the *Passim* analysis, and importing resulting *Passim* data into our internal web application for visualization and manual inspection. [46]

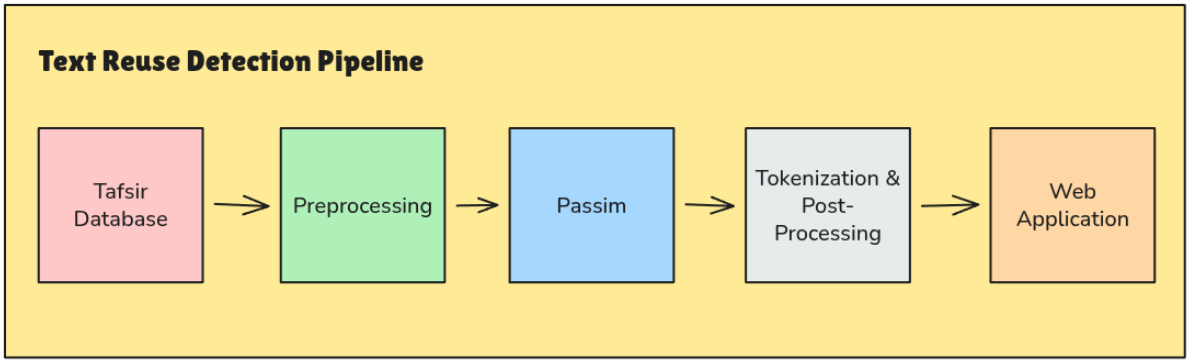


Figure 12: Processing pipeline for text-reuse detection using *Passim* on Tafsir subchapters commenting on Q 37:23. [Diagram: Thomas Jurczyk 2026]

We first extracted all 64 subchapters commenting on Q 37:23 from the Tafsir database. These subchapters often discuss neighboring verses, resulting in texts of varying lengths. To address this, we divided the files into 300-word chunks with a floating window of 150 words (see the Data section), yielding 285 text files. We also applied preprocessing steps, such as *hamza* normalization and the removal of vocalizations as well as of Qur'anic quotes. Finally, we converted the chunked subchapters – along with relevant metadata such as Tafsir titles and authors' death dates – into a JSONL (JSON Lines) file, the input format required by *Passim*. We used version 2.0.1 of *Passim* and executed the following basic command, to ensure that only later texts could be identified as drawing on earlier ones: [47]

```
passim --fields 'author_death_dce as year' --pairwise --filterpairs 'year < year2 AND uid<>uid2'
```
[48]

The date filtering also prevented individual chunks of a text from being compared with other chunks from the same text. Since we ran *Passim* multiple times with different parameter settings, we adjusted the required options (such as `-n` or `-u`) as needed. After manual review by our team and discussions with the *KITAB* team, the following parameter values yielded the best results for our data and research question. Our goal was not to identify all potential reused passages but to find interesting anchors for further close reading.[64] [49]

[64] The detailed parameter values that mostly include *Passim*'s default values are as follows: `minDF=2, maxDF=30, min_match=10, n=8, floating_ngrams=False, complete_lines=False, gap=200, max_offset=20, beam=20, pcopy=0.8, min_align=50, src_overlap=0.9, dst_overlap=0.5`.

	N-gram length (chars) (-n)	Gap parameter (chars) (-g)	Min. matching n-grams (-m)	Max. document frequency (-u)
best configuration	8	200	10	30

Table 4: Best configuration for using *Passim* with our Tafsir data.

This resulted in the following command: [50]

```
passim --fields 'author_death_dce as year' --pairwise --filterpairs 'year < year2 AND uid<>uid2' -n 8 -m 10 -g 200 -u 30
```
[51]

After running *Passim* with the specified parameter values, we tokenized the resulting alignments and clusters and processed them further to enable text search within our internal web application. The processed data were stored in a MongoDB database and integrated into the application. In the web application, team members can filter *Passim* clusters using keywords, review the clusters, explore a DAG visualization of each cluster, add comments, bookmark clusters, and automatically translate texts (cf. Fig. 13). [52]



Figure 13: Illustrative card from our internal web application showing one document from a *Passim* cluster. The application makes it possible to generate automatic translations (using Mistral AI), comments on documents, visualize clusters, and bookmark & download favorites. It is used as an internal tool to facilitate evaluation and joint work with the *Passim* data for all team members. Note that the attribution of Qu'ranic verses is part of ongoing research to automatically classify Qur'anic quotes using LLMs and other approaches. Thus, the attributions still contain errors and ambiguities. [Source: A01 Text Reuse Application]

The accessibility of the text reuse clusters via the web application allowed our researchers to easily access and work with the *Passim* data while analyzing individual clusters. This underscores the need for user-friendly abstraction layers when working in interdisciplinary teams with diverse technical backgrounds. The remainder of the article presents a proof-of-concept case study. While Sections 2–3 have focused on the computational workflow, Section 4 applies close philological reading to one of the resulting clusters. This combination illustrates the complementary roles of distant and close reading in the analysis of Tafsir traditions. [53]

4. Case Study: Metaphor Dynamics in a Large Tafsir Text-Reuse Cluster

Post-processing the *Passim* data yielded 159 text reuse clusters from the Tafsir NQQ corpus discussing verse Q 37:23. Our goal was to use text reuse detection as a heuristic tool to help identify relevant material for our research. Our objective was not to perform an exhaustive evaluation or optimize metrics such as F1 scores or recall;[65] rather, we aimed to assess whether text reuse detection could provide new insights into our data and metaphor-related research. [54]

Given our interest in the transmission chains of related exegetical discourses of the »road of Hellfire« (***ṣirāṭ al-jaḥīm***) metaphor and its development throughout the ages, we narrowed the clusters down to those primarily dealing with that metaphor. Because some subchapters of the Tafsirs focus on one verse, whereas most subchapters discuss a larger, though varying, range of verses before and after Q 37:23, some text-reuse clusters centered around the discussion of neighboring but not immediately ***ṣirāṭ***-related verses. The filtering option in our web application allowed us to identify clusters mentioning the term ***ṣirāṭ*** and related keywords, such as *hādī* (»guide«), *sabīl* (»way«), *ṭarīq* (»path«), and *jaḥīm* (»Hellfire«). This resulted in 16 clusters containing the term ***ṣirāṭ*** (some also contained alternative terms). In the following, we concentrated on the cluster combining both the largest number of documents and the highest frequency of the applied search terms.[66] [55]

4.1 Verse Q 37:23

Q 37:23 reads: {فَٱهْدُوهُمْ إِلَىٰ صِرَٰطِ ٱلْجَحِيمِ} (***fa-hdūhum ilā ṣirāṭi l-jaḥīm***), »guide them to the road of Hellfire«. In its Qur'anic context, God instructs servants, probably angels, to punish sinners on Judgment Day. This passage poses a major exegetical challenge. It is the only Qur'anic occurrence in which two otherwise positively connoted core metaphors, GUIDANCE (*hidāya*) and the STRAIGHT AND SOLID ROAD (***ṣirāṭ***),[67] are applied to the eschatological punishment of the condemned. The surrounding verses reinforce this scenario. In Q 37:22, God commands that »those who have committed wrongdoing« be gathered together with »their companions« (*wa-azwājahum*).[68] In Q 37:24, another command follows: [56]

{وَقِفُوهُمْ إِنَّهُمْ مَّسْـُٔولُونَ} (***wa-qifūhum innahum masʾūlūna***), »stop them; they will be questioned«. [57]

Q 37:25 then formulates the question addressed to them: [58]

[65] For this type of task, we would also need a pre-annotated dataset to verify the detection rate of our configurations, which is something that we do not possess.
[66] CHUNKED Q23:37v3-2026-04-27, Cluster ID 6. This is the title of the cluster in the Codeberg repository (file: tr-cluster.clusters.json), where the complete JSON data can be found.
[67] Cf. Abdelnour 2024; Seidel / Buessow 2025a.
[68] Literally »their spouses«.

{مَا لَكُمْ لَا تَنَاصَرُونَ} (*mā lakum lā tanāṣarūna*), »What is the matter with you that you do not help one another?« [59]

Exegetes thus need to explain how this inversion of the *ṣirāṭ* metaphor should be understood, how ›guidance‹ toward punishment can be interpreted, and how the sequence of gathering, stopping, and questioning should be imagined. A further complication is that Hadith traditions interpret the *ṣirāṭ* in this eschatological context as a bridge suspended over Hell, which is famously described as »thinner than a hair and sharper than a sword«.[69] [60]

4.2 Overview of the Text Reuse Cluster Structure

The selected *Passim* cluster is the largest cluster generated in chunked mode (cf. Fig. 14). It covers a long chronological span from the early 11th to the early 21st century and a wide geographical range, extending from Afghanistan and Iran in the East to Syria, Egypt, Yemen, the Maghreb, and Spain (al-Andalus) in the West.[70] [61]

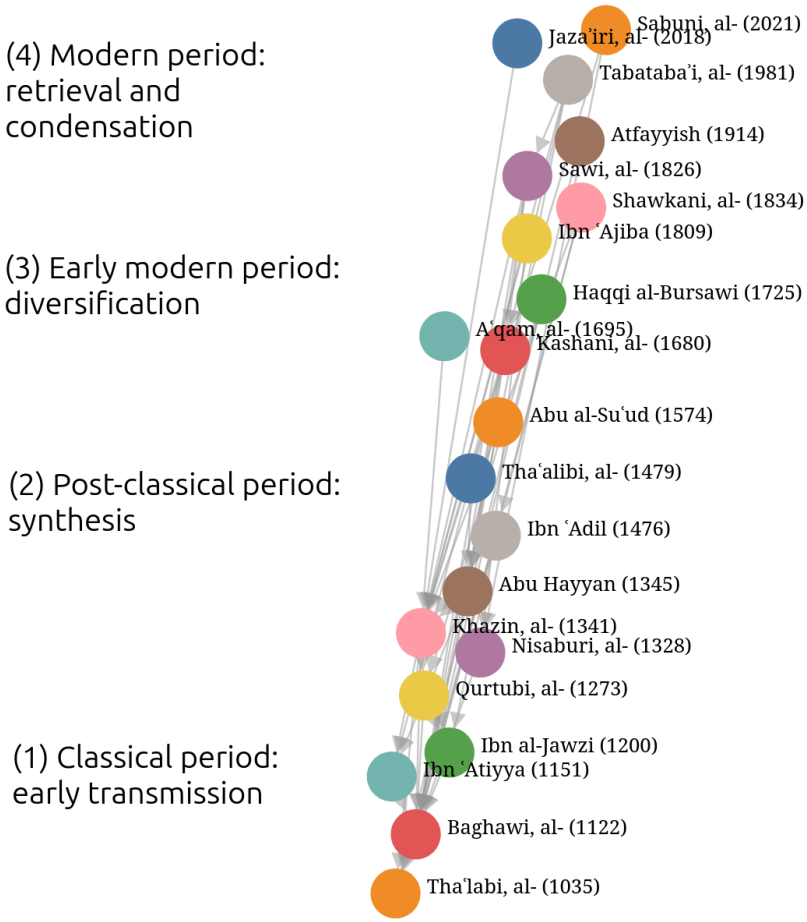


Figure 14: Chronological layers and exegetes in *Passim* cluster ID 6. Chronological directed acyclic graph (DAG) showing all authors and texts included in the cluster central to our analysis. Dates in parentheses indicate authors' death dates. Colors are assigned automatically and may be repeated because of the large number of texts represented. [Interactive graph: A01 Text Reuse Application, generated with D3.js]

[69] Al-Ḥajjāj 1955, Hadith no. 183.
[70] For a general background on the Tafsir works discussed in this article, cf. Shah / Abdel Haleem (eds.) 2020.

This cluster should not be understood as an independent exegetical tradition in its own right but as a heuristic tool that documents traces of repeated exegetical engagement with Q 37:23 and its surrounding verses. Its internal structure reveals four broad chronological layers:[71] (1) an early transmission spine centered on al-Thaʿlabi (d. 1035) and especially al-Baghawi (d. 1122), both from Central Asia, (2) a Mamluk and Andalusi phase of encyclopedic synthesis centered on the commentaries of the Syrian al-Khazin (d. 1341) and the Andalusian author Abu Hayyan (d. 1345), (3) an early modern phase of regional and confessional diversification, and (4) a modern phase characterized by selective retrieval, condensation, and recombination rather than the formation of new transmission hubs. In this sense, the cluster illustrates how shared exegetical material became stabilized over time while remaining open to changing ethical, eschatological, and doctrinal inflections. [62]

4.3 The Development of the *Ṣirāṭ al-Jaḥīm* Motif

This cluster of texts can be seen as an interconnected sub-corpus within the transtextual network of Qur'anic exegesis. It contains both intertextual relations, in the form of direct or indirect text reuse, and hypertextual relations, in that the later texts comment on, reinterpret, and expand earlier exegetical formulations. Viewed through the lens of metaphor dynamics, it provides one possible trace of how the Qur'anic expression ***ṣirāṭ al-jaḥīm*** was taken up and reconfigured in Tafsir. [63]

The Q 37:22–26 text-reuse cluster, centered on the expression »guide them to the road of Hellfire« (***fa-hdūhum ilā ṣirāṭi l-jaḥīm***) has developed over time from an early narrative and lexical exposition into an increasingly ethical, and doctrinally charged eschatological discussion. Later exegetes preserve a remarkable degree of textual continuity, especially through the dominant transmission line from al-Thaʿlabi to al-Baghawi and al-Khazin, as well as through related Mamluk-era syntheses. The designation Q 37:22–26 refers to the immediate narrative context of the exegetical discussion, whereas the text-reuse cluster itself was identified through passages relating to Q 37:23. Our analysis focuses primarily on Q 37:22–25, since Q 37:26 forms part of the same narrative scene but does not constitute a central element of the text-reuse patterns examined here. [64]

In earlier stages, the expression still retained the broader lexical meaning of a ›road‹ leading toward punishment. Later exegetes increasingly elaborated the scene as a sequence of detention, questioning, humiliation, doctrinal examination, and staged movement toward Hellfire. The result is not a single stable image, but a layered eschatological narrative in which the ***ṣirāṭ*** functions as a road, a route of coercive escort, a checkpoint, a threshold, or a bridge. Three prominent motifs emerged from the cluster: the moral grouping of the condemned, *hidāya* (guidance) as coerced escorting, and stopping and questioning as forms of eschatological control. [65]

4.4 The Moral Grouping of the Condemned

One of the most stable and widely reused motifs in the cluster is the moral grouping of sinners according to similarities and shared conduct. As early as in al-Thaʿlabi, this appears in the gloss on {وَأَزْوَاجَهُمْ} (*wa-azwājahum*), »and their companions«: [66]

[71] Our periodization combines contextual knowledge of Tafsir history with patterns visible in the reuse data, especially the emergence of transmission hubs, understood here as commentaries from which later authors repeatedly reused material. Al-Baghawi constitutes the most prominent hub in this cluster. In the two and a half centuries between al-Baghawi's commentary and Abu Hayyan's commentary, six commentators reused material from his treatment of Q 37:23. Over the following four centuries, only two further instances of reuse are attested, in the commentaries of Ibn ʿAdil and Kashani. This is followed by a phase of intensified renewed reuse, with four attestations within roughly 150 years, between the works of Shawkani and Tabatabaʾi. While this distribution does not by itself define historical periods, it supports our distinction between phases of dense transmission, encyclopedic synthesis, regional and confessional diversification, and selective retrieval. In this cluster, al-Baghawi appears foundational for the phases of early transmission and encyclopedic synthesis; his role is less visible during early modern regional and confessional diversification, before becoming relevant again in 19th- and 20th-century selective retrieval.

{فأهل الخمر مع أهل الخمر وأهل الزنا مع أهل الزنا} (*fa-ahl al-khamri maʿa ahl al-khamri wa-ahl al-zinā maʿa ahl al-zinā*), »The people of wine are with the people of wine, and the people of fornication with the people of fornication.« [67]

This rhetorically strong and polemical formula grouping people according to characteristic vices, becomes one of the central reuse nuclei of the cluster. Al-Baghawi preserved it almost verbatim, whereas later authors such as al-Nisaburi (d. 1328) and al-Khazin condensed and systematized it into a general principle of eschatological classification. In these later syntheses, the phrase no longer functions as transmitted narration but becomes part of a procedural order of resurrection: the condemned are assembled, classified , grouped, and then escorted toward punishment. [68]

The expansion in the commentary of the Mamluk Damascene scholar Ibn ʿAdil (d. 1476), who transformed the moral sorting formula into a confessional one is particularly striking: [69]

{واليهودي مع اليهودي والنصراني مع النصراني} (*wa-l-yahūdī maʿa l-yahūdī wa-l-naṣrānī maʿa l-naṣrānī*), »And the Jew with the Jew, and the Christian with the Christian.« [70]

This formula shifts the motif from moral resemblance toward confessional classification, suggesting a more comprehensive eschatological ordering of condemned groups. [71]

Closely related to this grouping motif is the recurring image that each unbeliever is paired or chained together with a devil. Al-Thaʿlabi transmitted the following tradition: [72]

{كل كافر مع شيطانه في سلسلة} (*kullu kāfirin maʿa shayṭānihi fī silsilatin*), »Every unbeliever is chained together with his devil.« [73]

This image was reused repeatedly by al-Baghawi, al-Qurtubi (d. 1273), al-Khazin, and others. The scene thus becomes increasingly processional: sinners are not only grouped by moral similarity, but are also physically bound to devils as personifications of their sins and arranged in ordered formations moving toward punishment. [74]

4.5 ›Guidance‹ (*Hidāya*) as Forced Escort

The second major node of reuse concerns the interpretation of »guide them« (*fa-hdūhum*). In the Qur'anic context, the verb normally evokes divine guidance, but throughout this cluster, exegetes have consistently tried to make sense of its inverse meaning by using linguistic arguments that the term is neutral and can thus be reinterpreted in coercive and processional terms. Al-Thaʿlabi listed several glosses: [75]

- {فادعوهم} (*fa-dʿūhum*), »Call / summon them.«
- {دلوهم} (*dallūhum*), »Direct them / show them the way.«
- {فدلوهم} (*fa-dallūhum*), »Then lead / direct them onward.«

A crucial lexical explanation follows immediately afterward: [76]

{والعرب تسمي السائق هادياً} »The Arabs call the driver or leader (*sāʾiq*) a ›guide‹ (*hādī*).« [77]

This explanation fundamentally reshapes the semantic field of *hidāya* in the passage. Guidance here is no longer salvific or spiritual. It means conducting, driving, escorting, or forcing onward toward punishment. The image is not of offering direction, but of compelled transfer. This lexical interpretation becomes one of the clearest identifiable reuse chains in the cluster. Al-Baghawi reproduced a closely related explanation, but used *al-sābiq* (»the one who precedes«) instead of ***al-sāʾiq*** (»the driver«), whereas later authors increasingly condensed the authorities and concentrated the material into anonymous scholastic prose. Al-Nisaburi preserved the full semantic triad: [78]

{ادعوهم أو قدموهم … أو دلوهم} »Call them, or bring them forward… or direct them.« [79]

The associated phrase, »The one who goes ahead is called the guide«, is also found in Ibn ʿAdil. Earlier texts used السائق (***al-sāʾiq***, »the driver conducting others«), which more directly evokes the sense of actively driving others towards a destination. By contrast, السابق (*al-sābiq*) means »the one who precedes« and does not necessarily imply guidance or escort. Ibn ʿAdil also transmits al-Wāḥidī's philological objection, stating *hādhā wahmun*, »this is a mistake«. However, this objection does not concern the designation of *al-sābiq* as *hādī*, but rather the interpretation of *hadā* as »bringing others forward« (*qaddama*) instead of »going ahead« (*taqaddama*). The variation between ***sāʾiq*** and *sābiq* thus highlights different aspects of the physical movement associated with *hidāya*, while al-Wāḥidī's objection introduces a further distinction between preceding others and causing them to move forward. Throughout the cluster, however, the dominant meaning remains remarkably stable: فاهدوهم (*fa-hdūhum*) signifies coercive movement toward Hellfire. [80]

4.6 The Detention at the *Ṣirāṭ* and the Questioning Scene

The most important development in the cluster concerns the relationship between ***ṣirāṭ al-jaḥīm***, detention, and interrogation. Earlier exegetes connected the questioning scene to a stopping point before entering Hellfire. The ***ṣirāṭ***, which elsewhere in the Qur'an evokes God's reliable guidance toward salvation, is inverted here: it does not lead toward righteousness but toward punishment. This inversion requires explanation. Al-Baghawi and Ibn al-Jawzi preserved the key formulation: [81]

{حبسوا عند الصراط لأن السؤال عند الصراط} (*ḥubisū ʿinda l-ṣirāṭi li-anna l-suʾāla ʿinda l-ṣirāṭi*), »They are detained at the ṣirāṭ because the questioning takes place at the ṣirāṭ.« [82]

This is one of the core transmission phrases of the cluster but the imagery remains deliberately indeterminate. The expression ***ʿinda l-ṣirāṭ*** could evoke the eschatological bridge over Hellfire, but the texts themselves do not necessarily describe a narrow or suspended bridge. Rather, the immediate image is that of a stopping point, checkpoint, or station of detention where interrogation occurs before punishment. [83]

The preposition ***ʿinda*** is significant. It locates the condemned ›at‹ the ***ṣirāṭ***, rather than ›on‹ or ›toward‹ it. This supports, without determining, the interpretation of the ***ṣirāṭ*** as a liminal site rather than simply as an extended road. One plausible hypothesis is that the bridge image is activated in part by this locative construction, especially when combined with the surrounding motifs of detention, questioning, and punishment. Lexically, ***ṣirāṭ*** means ›road‹, ›path‹, or ›way‹, often with connotations of straightness or firmness. Eschatologically, however, it can refer to the bridge stretched over Hellfire, an image developed in Hadith literature and later doctrinal tradition. The cluster thus maintains both possibilities: ***ṣirāṭ al-jaḥīm*** can be read as a road to Hellfire, a threshold before Hellfire, or a bridge-like locus of judgment. [84]

Ibn al-Jawzi, a late Abbasid Baghdadi moralist, intensified the procedural aspect of the scene by expanding the range of possible questions: [85]

{وفي هذا السؤال ستة أقوال} (*wa-fī hādhā l-suʾāli sittatu aqwālin*), »Regarding this questioning there are six opinions.« [86]

The questioning scene thus becomes increasingly elaborate and systematized. It is no longer only a brief moment of detention but a structured episode within the eschatological sequence. [87]

A major interpretive shift occurs in al-Qurtubi who described the narrative tension between the command: [88]

{فَاهْدُوهُمْ إِلَىٰ صِرَاطِ الْجَحِيمِ} (*fa-hdūhum ilā ṣirāṭi l-jaḥīmi*), »Guide them to the road of Hellfire«, [89]

and the following command: [90]

{وَقِفُوهُمْ إِنَّهُمْ مَسْئُولُونَ} (*wa-qifūhum innahum masʾūlūna*), »Stop them; indeed, they are to be questioned.« [91]

If the condemned are already being led to Hellfire, when does the questioning occur? Al-Qurtubi resolves this through rhetorical rearrangement: [92]

{وفيه تقديم وتأخير} (*wa-fīhi taqdīmun wa-taʾkhīrun*), »There is inversion [or rearrangement] in the order.« [93]

He explains: [94]

{قفوهم للحساب ثم سوقوهم إلى النار} (*qifūhum li-l-ḥisābi thumma sūqūhum ilā l-nāri*), »Detain them for reckoning, then drive them to the Fire.« [95]

This is crucial for the metaphor dynamics of the passage. Rather than imagining the questioning simply at a bridge-like crossing, al-Qurtubi reconstructs a procedural sequence that consists of detention, reckoning, and only afterward forced movement into Hellfire. The eschatological scene becomes increasingly administrative and juridical. [96]

Al-Nisaburi developed this procedural imagery further. After reusing the earlier lexical discussion of *fa-hdūhum* (»lead them«), he adds: [97]

{إلى صراط الجحيم وسطها} (*ilā ṣirāṭi l-jaḥīmi wasaṭahā*), »to the road of Hellfire, into its midst«. [98]

This redirects the imagery away from the bridge as a passage over Hellfire. The *ṣirāṭ al-jaḥīm* becomes a route leading into Hellfire itself. The accompanying verbs reinforce the processional structure: [99]

- {احشروا} (*uḥshurū*), »gather them«.
- {قدموهم} (*qaddimūhum*), »bring them forward«.
- {دلوهم} (*dallūhum*), »direct them«.
- {احبسوهم} (*iḥbisūhum*), »detain them«.

The eschatological scene now resembles an organized movement of condemned groups through successive stages of classifying, escorting, detention, humiliation, and final delivery to punishment. [100]

Abu Hayyan introduced another important shift by linking the verse explicitly to Q 6:27: [101]

{وَلَوْ تَرَىٰ إِذْ وُقِفُوا عَلَى النَّارِ} (*wa-law tarā idh wuqifū ʿalā l-nāri*), »If only you could see when they are made to stand before the Fire.« [102]

This again transformed the scene from juridical detention to psychological exposure. The condemned are halted before Hellfire itself and are forced to behold it. The questioning is therefore not merely procedural; it becomes humiliating and demonstrative. [103]

A further major reuse chain running through the cluster is the Hadith: [104]

{لا تزول قدما عبد يوم القيامة حتى يسأل} (*lā tazūlu qadamā ʿabdin yawma l-qiyāmati ḥattā yusʾala*), »On the Day of Judgment, a servant's feet will not move until he is questioned.«[72] [105]

This statement functions as a hermeneutical anchor for interpreting *wa-qifūhum innahum masʾūlūna*, »Stop them; indeed, they are to be questioned.« In earlier layers, it primarily universalized accountability: all human beings will be detained and interrogated. Over time, however, the number and content of the questions change. In al-Baghawi, each human being is questioned on four issues: [106]

{عن شبابه فيما أبلاه، وعن عمره فيما أفناه، وعن ماله} (*ʿan shabābihi fīmā ablāhu, wa-ʿan ʿumrihi fīmā afnāhu, wa-ʿan mālihi*), »about his youth and how he wore it out, about his life and how he spent it, and about his wealth«. [107]

Ibn ʿAtiyya expanded the formulation by adding knowledge as an object of accountability: [108]

{وعما عمل فيما علم} (*wa-ʿammā ʿamila fīmā ʿalima*), »and about what he did with the knowledge he possessed«. [109]

The movement from four to five questions not only reflects numerical variation but ethical expansion where knowledge itself becomes an object of eschatological examination. [110]

Later exegetes continued to reuse and adapt the report. Al-Khazin integrated it into a harmonized eschatological sequence of gathering, escort, detention, interrogation, and rebuke. In al-Kashani's Shiʿi reinterpretation, however, the Hadith underwent a decisive doctrinal transformation: alongside questions about youth, lifespan, and wealth, the interrogation is extended to specifically Shiʿi matters of belief and allegiance. [111]

Thus, in these later commentaries, physical movement toward Hellfire recedes, while the theological content of interrogation becomes central. [112]

One of the most sophisticated developments within the cluster occurs in Ibn ʿAtiyya. Earlier exegetes generally understood the questioning in universal or doctrinal terms, as accountability for misdeeds or lack of belief. Ibn ʿAtiyya, however, linked the interrogation directly to the following verse: [113]

{مَا لَكُمْ لَا تَنَاصَرُونَ} (*mā lakum lā tanāṣarūna*), »What is with you that you do not help one another?« [114]

He explained: [115]

[72] Al-Nawawi 2000, no. 407, p. 93; al-Tirmidhi 1996, no. 2417.

{أي إنكم مسؤولون عن امتناعهم عن التناصر} (*ay innakum masʾūlūna ʿan imtināʿihim ʿan al-tanāṣuri*), »That is, you are being held accountable for refraining from supporting one another.« [116]

This introduces a new ethical dimension into the cluster where the questioning exposes the collapse of worldly solidarities in those who are condemned. Individuals who once reinforced one another in unbelief, oppression, or arrogance are now unable to help one another. The worldly boast of collective strength is thus inverted into eschatological helplessness. [117]

This is closely related to the motif of ridicule. Ibn ʿAtiyya preserved a comment by Ibn Masʿud: [118]

{يسألون هل يحبون شرب الماء البارد} (*yusʾalūna hal yuḥibbūna shurba l-māʾi l-bāridi*), »They will be asked whether they liked drinking cold water.« [119]

He immediately clarified: [120]

{وهذا على طريق الهزء بهم} (*wa-hādhā ʿalā ṭarīqi l-hazʾi bihim*), »This is by way of mockery toward them.« [121]

The questioning scene thus not only becomes juridical, but also humiliating and mockingly ironic, which intensifies its affective dimension. God does not ask for information; rather the interrogation publicly exposes the condemned and intensifies their disgrace. This mocking becomes even stronger in Abu Hayyan, where the sinners are halted before the Fire itself and made to behold it. [122]

Taken together, these developments show that the *ṣirāṭ al-jaḥīm* cluster does not stabilize a single image of the *ṣirāṭ*. Rather, it transforms the Qur'anic »road of Hellfire« into a broader symbolic zone of coercive movement, detention, exposure, questioning, humiliation, and doctrinal reckoning. The *ṣirāṭ* can function as road, route, checkpoint, threshold, or bridge-like site of judgment, depending on how the reused material is arranged and interpreted. [123]

4.7 Discussion

This analysis suggests that the al-Thaʿlabi–al-Baghawi reuse cluster represents a highly stable yet internally dynamic exegetical formation. Through interpretation and repeated reuse, the relatively brief Qur'anic phrases from Q 37:22–26 generate an increasingly elaborate eschatological narrative. The cluster begins with lexical and narrative glosses concerning grouping, escort, detention, questioning, and punishment, but gradually develops into a more complex procedural, moralized and emotionalized scene of Judgment Day. Its stability lies in the persistence of shared explanatory elements; its dynamism lies in the ways these elements are condensed, rearranged, expanded, and doctrinally reoriented by later exegetes. [124]

The shared concern of all exegetical texts in the cluster is how to transform the compressed Qur'anic sequence of gathering, guidance, stopping, and questioning into a coherent eschatological scenario. The reused material provides these exegetes with a stable toolkit for explaining the inversion of *hidāya*, the moral grouping of the condemned, and the location and purpose of interrogation at the *ṣirāṭ*. Text reuse thus serves to stabilize an interpretive solution while leaving room for later moral, procedural, and doctrinal reconfiguration. [125]

One of the key findings of the clustering procedure concerns the ambiguity of ***ṣirāṭ al-jaḥīm***. It showed that the cluster does not consistently stabilize the expression as the later theological »bridge over Hellfire«, even though bridge-like associations are evoked in some passages. Rather, the dominant imagery is processional and custodial: sinners are gathered, grouped according to their conduct or similarities, chained with devils, escorted toward punishment, detained for questioning, exposed to humiliation, and finally delivered to Hellfire itself. The recurrent phrase {حبسوا عند الصراط لأن السؤال عند الصراط} (***ḥubisū ʿinda l-ṣirāṭi li-anna l-suʾāla ʿinda l-ṣirāṭi***), »They are detained at the ṣirāṭ because the questioning takes place at the ṣirāṭ«, permits a bridge-like reading, but does not require it. The ***ṣirāṭ*** functions as a flexible eschatological site as a road, a route of forced escort, a stopping point, a checkpoint, a threshold, or a bridge-like locus of judgment. [126]

The reuse chain also demonstrates that textual continuity does not imply interpretive stasis. Later exegetes from the post-classical period onward (chronological layers 2, 3, and 4, cf. Fig. 14) repeatedly reconfigure inherited material while continuing to build on shared formulations. Ibn ʿAtiyya (d. 1151) moralized the questioning by linking it to the collapse of solidarity among the condemned; al-Qurtubi proceduralized the order of events by separating detention, reckoning, and subsequent driving to the Fire, Abu Hayyan intensified the scene as one of psychological exposure before Hellfire, and the Safavid-era commentary from the early modern period (chronological layer 3) by al-Kashani (d. 1680) reoriented the interrogation toward specifically Shiʿi concerns with religious authority (*walāya*). These examples show that text reuse can preserve an exegetical resource while simultaneously enabling new ethical, eschatological, and doctrinal emphases. [127]

The later chronological layers 3 and 4 of the cluster confirm this pattern of continuity and reconfiguration. Ottoman, Shiʿi, and modern Sunni exegetes did not create an entirely new interpretive tradition; rather, they retrieved and reorganized inherited elements in ways that reflect changing pedagogical, ethical, and doctrinal priorities. The Ottoman-era commentary by al-Bursawi (d. 1725) preserves and expands the moral and confessional grouping formula while intensifying the motif of humiliating exposure. The modern Shiʿi philosopher and exegete al-Tabatabaʾi (d. 1981), building on earlier Shiʿi trajectories, transforms the universal moral interrogation into a doctrinal examination. Modern Sunni commentaries, such as those by al-Jazaʾiri (d. 2018) and al-Sabuni (d. 2021), reduce the material to compact pedagogical summaries while preserving its main narrative nodes: grouping, coerced guidance, detention ›at‹ the ***ṣirāṭ***, interrogation, and rebuke. In this late phase, inherited motifs function less as sources of new narrative elaboration than as stable elements for selection, compression, and didactic redeployment. [128]

This case study thus demonstrates how shared exegetical material has become stable over time. The close reading reveals how this material has been condensed, rearranged, and semantically altered. The explored cluster thus represents a case in which philological and narrative explanations became a shared exegetical resource: the metaphor of the »road of Hellfire« (***ṣirāṭ al-jaḥīm***) was not fixed once and for all, but has repeatedly been reconfigured within shifting eschatological, ethical, and doctrinal frameworks. [129]

5. Conclusion

This article has shown how text reuse detection with *Passim* can support the study of intertextuality and metaphor dynamics in Tafsir literature. Rather than using *Passim* to identify all instances of reuse between documents, we treated it as a heuristic instrument for locating meaningful sites of exegetical continuity, condensation, and reconfiguration. The case study on Q 37:23 demonstrates that computational text reuse detection can help identify a shared exegetical resource that developed across time, regions, and confessional contexts, while close reading remains indispensable for interpreting what this reuse means. Consequently, it serves as a natural complement to our previous text-mining study, which we used to [130]

identify broader patterns in our Tafsir data and to pinpoint the discussion of Q 37:23 as a compelling topic for examining WAY metaphors. In this study, we expand on that foundation by adding a chronological and intertextual focus.

The selected *Passim* cluster shows how a compressed Qur'anic scene was expanded into a complex eschatological narrative. The reused material helped stabilize a set of recurring exegetical motifs: the moral grouping of the condemned, *hidāya* as coerced escorting, and stopping and questioning. At the same time, the cluster also reveals significant interpretive flexibility. Later exegetes did not merely transmit earlier material; they condensed it, rearranged it, moralized it, proceduralized it, and in some cases reoriented it doctrinally. The metaphor of *ṣirāṭ al-jaḥīm* was therefore not fixed once and for all. It remained semantically productive, oscillating between road, route of forced movement, threshold, checkpoint, and bridge-like site of judgment. [131]

This makes the case study a proof of concept for a broader methodological claim. *Passim* and text reuse detection in general are particularly useful when it is not expected to replace philological interpretation, but to guide it. It allows researchers to detect patterns that would be difficult to identify through close reading alone, especially in a large and internally diverse corpus. Yet the intellectual-historical significance of these patterns only becomes visible when the computational results are interpreted in relation to genre, chronology, doctrinal orientation, and metaphorical structure. In this sense, text reuse detection provides a bridge between distant reading and close reading: it points to textual constellations that require further analysis and helps make the selection of case studies more transparent. [132]

Several methodological limitations remain. The present study relied on a sliding-window approach with 300-word chunks in order to handle Tafsir texts of very different length and structure. This approach proved effective for discovery, but it also produced overlapping duplicates that inflated some clusters. Future work should adopt more refined procedures, such as those developed in the *KITAB* context, that can trace reuse beyond chunk boundaries without duplicating text segments.[73] It would also be valuable to test *alNaql*, a tool designed specifically for Arabic text reuse detection, once it becomes suitable for larger multi-document comparisons. [133]

Further improvements concern the integration of text reuse detection into the Tafsir web application. Tafsir subchapters differ greatly in length and scope; some comment on a single verse, while others cover longer passages. We therefore plan to add a functionality that retrieves and merges subchapters across a specified verse range, for example Q 37:20–26[74]. This range follows the division into coherent passages adopted by the ›Corpus Coranicum‹ project. Q 37:20–21 introduce the eschatological scene of Judgment Day, which unfolds in the subsequent verses and includes the gathering of the condemned, their guidance toward Hellfire, their detention and questioning, and their eventual submission. In addition, more precise classification of Qur'anic quotations would allow us to generate customized text versions, such as removing quotations from the verses under discussion while retaining cross-references to other Qur'anic passages. [134]

Finally, future work should compare the cluster discussed here with other types of *Passim* clusters. One promising comparison is a more theologically charged cluster centered on Abu l-Layth al-Samarqandi and Muhammad ibn Hasan al-Tusi,[75] where reuse appears to engage more directly with doctrinal argumentation. Such comparisons would allow us to move from a single proof of concept toward a broader typology of text [135]

[73] Cf. Barber 2020.
[74] See Hartwig / Neuwirth 2026, commentary on Sura 37, section Verse 20–26 and the sura outline, which identifies Q 37:20–26 as a coherent textual unit under the heading »Eschatologische Szenerie, Vorgang, Weheruf der Gerichteten«.
[75] Name of the cluster in the cluster data (see Codeberg repository): CHUNKED Q23:37v3-2026-04-27, Cluster ID 0.

reuse and metaphor dynamics in Tafsir literature. This would also clarify how text mining and text reuse analysis can complement one another as part of a larger methodological toolbox for the study of religious meaning-making.

Transparency Note on the Use of AI in this Article

[136] The Mistral AI chat interface was used to revise the original English paragraphs in this article with the prompt: »Please revise the following English texts. Stick as closely as possible to the original text.« Mistral AI's Saba model was also used to generate basic translations of Arabic texts in the project's web application. These translations were particularly helpful for the digital humanities researcher co-authoring this article, who does not read the source languages. In addition, ChatGPT was used during the drafting and revision process to improve clarity, structure, concision, English style, and consistency of terminology and transliteration. It was also used to examine the article's argumentation, case-study structure, and the formulation of methodological and interpretive claims. All AI-generated suggestions were reviewed, evaluated, and revised by the authors.

Acknowledgments

[137] This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1475 – Project ID 441126958. The authors would like to thank Tatjana Scheffler and Christoph Sander for their insightful comments on an earlier draft of this study and Esther Singer for her careful copy-editing.

Bibliography

Mohammed Gamal Abdelnour: The Straight Path (*al-Ṣirāṭ al-Mustaqīm*). In: Brendan N. Wolfe et al. (eds.): St Andrews Encyclopaedia of Theology. St Andrews, UK 2024. [online]

Graham Allen: Intertextuality. In: Deidre Shauna Lynch (ed.): Oxford Research Encyclopedia of Literature. Oxford 2016–2026. 23.05.2019. DOI: 10.1093/acrefore/9780190201098.013.1072

Mathew Barber: Adventures in Alignments: Training an Algorithm to Recognise Text Reuse. In: KITAB. 07.08.2020. HTML. [online]

Adrian Bernhard: Development of the Quranic WAY Metaphor in Tafsir: A Corpus Analysis Approach. Master's thesis, Ruhr-Universität Bochum. 2024. [GVK record]

Oran Brill / Moshe Koppel / Avi Shmidman: FAST: Fast and Accurate Synoptic Texts. In: Digital Scholarship in the Humanities 35 (2020), no. 2, pp. 254–264. 23.05.2019. DOI: 10.1093/llc/fqz029

Frank Buschmann / Aron Fink: Dotplots und Alignments als Untersuchungsinstrumente für Überlieferungsgemeinschaften am Beispiel von Priameln. In: Zeitschrift für digitale Geisteswissenschaften 8 (2023). 14.12.2023. DOI: 10.17175/2023_009

Paul Clough / Robert Gaizauskas / Scott S. L. Piao / Yorick Wilks: METER: MEasuring TExt Reuse. In: 40th Annual Meeting of the Association for Computational Linguistics. Proceedings of the Conference. (ACL '02, Philadelphia, 07.–12.07.2002). Kerrville, US-TX 2002, pp. 152–159. DOI: 10.3115/1073083.1073110

Neil Coffee / Jean-Pierre Koenig / Shakthi Poornima / Christopher W. Forstall / Roelant Ossewaarde / Sarah L. Jacobson: The Tesserae Project: Intertextual Analysis of Latin Poetry. In: Literary and Linguistic Computing 28 (2013), no. 2, pp. 221–228. 20.07.2013. DOI: 10.1093/llc/fqs033

Ryan Cordell / David Smith: The Viral Texts Project. Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines. Last updated: 2024. HTML. [online]

Marten Düring / Matteo Romanello / Maud Ehrmann / Kaspar Beelen / Daniele Guido / Brecht Deseure / Estelle Bunout / Jana Keck / Petros Apostolopoulos: *impresso* Text Reuse at Scale. An Interface for the Exploration of Text Reuse Data in Semantically Enriched Historical Newspapers. In: Frontiers in Big Data 6 (2023). 03.11.2023. DOI: 10.3389/fdata.2023.1249469

Greta Franzini / Marco Passarotti / Maria Moritz / Marco Büchler: Using and Evaluating TRACER for an *Index Fontium Computatus* of the *Summa contra Gentiles* of Thomas Aquinas. In: Elena Cabrio / Alessandro Mazzei / Fabio Tamburini (eds.): Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018, Turin, 10.–12.12.2018). Bonn 2018, pp. 200–206. PDF. [online]

Henning Gebhard / Vandana Jha / Philipp Tögel / Stefanie Dipper / Frederik Elwert / Danah Tonne: Metaphors of Religion. Towards a Shared Infrastructure for Metaphor Analysis. Poster, DHd2023: Open Humanities, Open Culture. 9. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum, Trier / Belval, LU, 13.–17.03.2023. Zenodo. 10.03.2023. Version v2: 10.03.2023. DOI: 10.5281/zenodo.7711495

Gérard Genette: Palimpsests. Literature in the Second Degree (= Stages, 8). Lincoln, US-NE 1997. [GVK record]

Claude Gilliot: Exegesis of the Qur'an: Classical and Medieval. In: Jane Dammen McAuliffe (ed.): Encyclopaedia of the Qur'an. EQ. 5 volumes. Leiden 2001–2006. Volume 2 (2002): E–I, pp. 99–124. [GVK record]

Andreas Görke / Johanna Pink (eds.): Tafsīr and Islamic Intellectual History: Exploring the Boundaries of a Genre (= Qur'anic Studies Series, 12). Oxford etc. 2014. [GVK record]

Ronald Haentjens Dekker / Dirk van Hulle / Gregor Middell / Vincent Neyt / Joris van Zundert: Computer-Supported Collation of Modern Manuscripts: CollateX and the Beckett Digital Manuscript Project. In: Digital Scholarship in the Humanities 30 (2015), no. 3, pp. 452–470. 19.04.2014. DOI: 10.1093/llc/fqu007

Muslim ibn al-Ḥajjāj: Ṣaḥīḥ Muslim. Cairo 1955. [GVK record]

Dirk Hartwig / Angelika Neuwirth, with contributions by Ali Aghaei and Tolou Khademalsharieh: Sure 37 – *aṣ-ṣāffāt* – »Die in Reihen Schreitenden«. In: Berlin-Brandenburgische Akademie der Wissenschaften (ed.): Corpus Coranicum. Last accessed: 13.11.2026. HTML. [online]

Stefan Jänicke / Annette Geßner / Marco Büchler / Gerik Scheuermann: Visualizations for Text Re-Use. In: Robert S. Laramee / Andreas Kerren / José Braz (eds.): Proceedings of the 5th International Conference on Information Visualization Theory and Applications (VISIGRAPP, Lisbon, 05.–08.01.2014). Setúbal, PT 2014, pp. 59–70. DOI: 10.5220/0004692500590070

Thomas Jurczyk / Roman Seidel / Adrian Bernhard / Tatjana Scheffler / Johann Buessow: Text Mining Tafsir: Compilation and Preliminary Explorations of a Curated Corpus of 80 Qur'anic Commentaries. In: Journal of Digital Islamicate Research 3 (2025), no. 1, pp. 97–167. DOI: 10.1163/27732363-bja00010

Tynan Kelly: Detecting Text Reuse in Historical Arabic Texts: Challenges and Strategies. In: Journal of Digital Islamicate Research 3 (2025), no. 2, pp. 262–295. 14.11.2025. [online]

Volkhard Krech / Tim Karis / Frederik Elwert: Metaphors of Religion: A Conceptual Framework (= Metaphor Papers, 1). 2023. [online]

Vladimir Iosifovich Levenshtein: Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. In: Soviet Physics. Doklady 10 (1966), no. 8, pp. 707–710. [GVK record]

Hadar Miller / Tsvi Kuflik / Moshe Lavee: Text Alignment in the Service of Text Reuse Detection. In: Applied Sciences 15 (2025), no. 6. 20.03.2025. DOI: 10.3390/app15063395

Yahya ibn Sharaf al-Nawawi: Riyāḍ al-ṣāliḥīn min kalām sayyid al-mursalīn. 3rd edition. Beirut 2000. [online]

Saul Ben Needleman / Christian D. Wunsch: A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. In: Journal of Molecular Biology 48 (1970), no. 3, pp. 443–453. 28.03.1970. DOI: 10.1016/0022-2836(70)90057-4

Lorenz Nigst / Maxim Romanov / Sarah Bowen Savant / Masoumeh Seydi / Peter Verkinderen / Hamidreza Hakimi: OpenITI: A Machine-Readable Corpus of Islamicate Texts. Zenodo. 20.05.2019. Version 2023.1.8: 17.10.2023. Dataset. DOI: 10.5281/zenodo.10007820

Ossama Obeid / Nasser Zalmout / Salam Khalifa / Dima Taji / Mai Oudah / Bashar Alhafni / Go Inoue / Fadhl Eryani / Alexander Erdmann / Nizar Habash: CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing. In: Nicoletta Calzolari / Frédéric Béchet / Philippe Blache / Khalid Choukri / Christopher Cieri / Thierry Declerck / Sara Goggi / Hitoshi Isahara / Bente Maegaard / Joseph Mariani / Hélène Mazo / Asuncion Moreno / Jan Odijk / Stelios Piperidis (eds.): Proceedings of the Twelfth Language Resources and Evaluation Conference (Marseille, 11.–16.05.2020). Paris 2020, pp. 7022–7032. Version 2. [online]

Johanna Pink: Tradition, Authority and Innovation in Contemporary Sunnī *tafsīr*: Towards a Typology of Qur'an Commentaries from the Arab World, Indonesia and Turkey. In: Journal of Qur'anic Studies 12 (2010), no. 1–2, pp. 56–82. DOI: 10.3366/jqs.2010.0105

Johanna Pink: *Tafsīr* as Discourse. Institutions, Norms, and Authority. In: Majid Daneshgar / Aaron W. Hughes (eds.): Deconstructing Islamic Studies (= The Mizan Series, 4). Boston etc. 2020, pp. 53–76. [GVK record]

Andrew Rippin: Tafsīr. In: Peri J. Bearman / Thierry Bianquis / Clifford Edmund Bosworth / Emericus van Donzel / Wolfhart Peter Heinrichs (eds.): The Encyclopaedia of Islam. New edition. 12 volumes. Leiden etc. 1954–2009. Volume 10 (2000): T–U, pp. 83–88. [GVK record]

Matteo Romanello / Simon Hengchen: Detecting Text Reuse with Passim. In: Programming Historian. 16.05.2021. HTML. [online]

Samuel J. Ross: What Were the Most Popular *tafsīr*s in Islamic History? Part 1: An Assessment of the Manuscript Record and the State of *tafsīr* Studies. In: Journal of Qur'anic Studies 25 (2023), no. 3, pp. 1–54. DOI: 10.3366/jqs.2023.0555

Roman Seidel: Qur'anic Lead Metaphors and the Unfolding of their Religious Meaning: Considerations on the Theory of Metaphor and Transtextuality Applied to Qur'an and Its Exegesis. In: Asiatische Studien 80 (2026), no. 2, pp. 357–429. DOI: 10.1515/asia-2025-0060

Roman Seidel / Johann Buessow (2025a): Mapping the ›Straight and Solid Road‹ within the Qur'anic Landscape: Development of a Primordial Islamic Way Metaphor. In: Kianoosh Rezania / Neda Mohtashami / Roman Seidel (eds.): Conceptualizing CONDUCT OF LIFE through WAY Metaphors (= Metaphor Papers, 19). Bochum 2025. DOI: 10.46586/mp.430.507

Roman Seidel / Johann Buessow (2025b): Annotating Explicative Metaphors: An Annotation Guideline for Qur'anic Lead Metaphors and Explicative Metaphors in Exegetical Texts (Tafsir) (= Metaphor Papers, 31). Bochum 2025. DOI: 10.46586/mp.451

Mustafa Akram Ali Shah / Muhammad Abdel Haleem (eds.): The Oxford Handbook of Qur'anic Studies (= Oxford Handbooks). Oxford 2020. DOI: 10.1093/oxfordhb/9780199698646.001.0001

Nicolai Sinai: The Qur'an. A Historical-Critical Introduction. Edinburgh 2017. DOI: 10.1515/9780748695782

David A. Smith: Passim. GitHub. 17.09.2015. Version 2.0.1: 07.02.2024. [online]

David Smith / Ryan Cordell / Abby Mullen: Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers. In: Ryan Cordell / David Smith: Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines. 22.05.2015. HTML. [online]

Temple Ferris Smith / Michael Spencer Waterman: Identification of Common Molecular Subsequences. In: Journal of Molecular Biology 147 (1981), no. 1, pp. 195–197. 25.03.1981. DOI: 10.1016/0022-2836(81)90087-5

Donald Sturgeon: Digital Approaches to Text Reuse in the Early Chinese Corpus. In: Journal of Chinese Literature and Culture 5 (2018), no. 2, pp. 186–213. 01.11.2018. DOI: 10.1215/23290048-7256963

Muhammad ibn ʿIsa al-Tirmidhi: al-Jāmiʿ al-kabīr (Sunan al-Tirmidhī). Beirut 1996. [WorldCat record]

Tariq Yousef / Stefan Jänicke: A Survey of Text Alignment Visualization. In: IEEE Transactions on Visualization and Computer Graphics 27 (2021), no. 2, pp. 1149–1159. 12.10.2020. DOI: 10.1109/TVCG.2020.3028975

Aleksi Vesanto / Asko Nivala / Heli Rantala / Tapio Salakoski / Hannu Salmi / Filip Ginter: Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771-1910. In: Gerlof Bouma / Yvonne Adesam (eds.): Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language (Gothenburg, 22.05.2017). Linköping 2017, pp. 54–58. [online]

List of Figures and Tables

Figure 1: Genesis of the Tafsir corpus. [Diagram: Thomas Jurczyk 2026]

Figure 2: Tafsirs by century (according to their authors' assumed death dates). [Source: Bernhard 2024, p. 81]

Figure 3: Tafsirs by region. [Source: Bernhard 2024, p. 82]

Figure 4: Screenshot of the A01 Tafsir application, showing an overview of Tafsir texts with metadata annotations. [Source: Tafsir app]

Figure 5: *KITAB* visualization of one-to-one text reuse in ʿAntara Ibn Shaddād, *Diwan* and Luwīs Shīkhū al-Yasūʿī, *Shuʿarāʾ al-naṣrāniyya baʿda l-islām*. [Source: KITAB]

Figure 6: *KITAB* corpus-wide one-to-many visualization of Abu Talib ʿAbd Manaf, *Diwan*. [Source: KITAB]

Figure 7: Chronological text reuse in the form of an acyclic directed graph (DAG). [Source: Graph – Text Reuse in Rare Books]

Figure 8: Exemplary DAG visualization of a *Passim* cluster on our internal text-reuse detection website. [Interactive graph: A01 Text Reuse Application, generated with D3.js]

Table 1: Illustrative matrix-based text alignment table.

Figure 9: Variant graph for the two documents W1 (»The quick red fox.«) and W2 (»The quick brown fox.«) created using the *CollateX* web interface. [Source: CollateX]

Figure 10: Scatter plot and boxplot of the text lengths (characters) of the preprocessed Corpus NQQ. This corpus was stripped of Qu'ranic quotes and the preprocessing pipeline was applied (including removal of punctuation marks and non-Arabic characters). Whitespaces were also ignored. [Plot: Thomas Jurczyk 2026]

Table 2: Descriptive statistics for the two subchapter corpora. The corpus without Qur'anic quotes was stripped of Qur'anic quotes and the preprocessing pipeline was applied (including removal of punctuation marks and non-Arabic characters). Whitespaces were also ignored.

Figure 11: Scatter plot and boxplot of the text lengths (characters) of the preprocessed and chunked Corpus NQQ. This corpus was stripped of Qu'ranic quotes and the preprocessing pipeline was applied (including removal of punctuation marks and non-Arabic characters). Whitespaces were also ignored. [Plot: Thomas Jurczyk 2026]

Table 3: Descriptive statistics for the two subchapter corpora with chunking applied. The corpus without Qur'anic quotes was stripped of Qur'anic quotes and the preprocessing pipeline was applied (including removal of punctuation marks and non-Arabic characters). Whitespaces were also ignored.

Figure 12: Processing pipeline for text-reuse detection using *Passim* on Tafsir subchapters commenting on Q 37:23. [Diagram: Thomas Jurczyk 2026]

Table 4: Best configuration for using *Passim* with our Tafsir data.

Figure 13: Illustrative card from our internal web application showing one document from a *Passim* cluster. The application makes it possible to generate automatic translations (using Mistral AI), comments on documents, visualize clusters, and bookmark & download favorites. It is used as an internal tool to facilitate evaluation and joint work with the *Passim* data for all team members. Note that the attribution of Qu'ranic verses is part of ongoing research to automatically classify Qur'anic quotes using LLMs and other approaches. Thus, the attributions still contain errors and ambiguities. [Source: A01 Text Reuse Application]

Figure 14: Chronological layers and exegetes in *Passim* cluster ID 6. Chronological directed acyclic graph (DAG) showing all authors and texts included in the cluster central to our analysis. Dates in parentheses indicate authors' death dates. Colors are assigned automatically and may be repeated because of the large number of texts represented. [Interactive graph: A01 Text Reuse Application, generated with D3.js]