Bridging Accessibility and Rigour: A Web App for Authorship Analysis Incorporating KWIC Exploration and Shared N-gram Profiling

John Blake and Krzysztof Kredens

Aston Institute for Forensic Linguistics, Aston University, Birmingham,United Kingdom.

*Corresponding author(s). E-mail(s): j.blake3@aston.ac.jp; k.j.kredens@aston.ac.uk;

Abstract

The increasing volume of digital linguistic evidence and the expansion of virtual communicative spaces demand forensic linguistic tools that are both methodologically rigorous and accessible to a wider range of users, including legal practitioners. This paper introduces a bespoke, web-based authorship analysis tool designed to identify and visualize key textual patterns across questioned and known datasets. Operating through an intuitive graphical user interface, the tool offers an accessible alternative to command-line methods, expanding the reach of computational forensic linguistics to non-technical users. The tool integrates two core functionalities: (1) automated identification and comparison of shared n-grams, particularly bigrams and trigrams, within and between datasets, and (2) Key Word in Context (KWIC) exploration to support the interpretation of overlapping or distinctive lexical patterns. By combining these capabilities, the system enables both confirmatory and exploratory analysis of linguistic evidence, supporting investigations where authorship is contested or where stylistic consistency must be assessed. It has applications across the three domains of forensic linguistics: language as evidence, interaction in legal settings, and language of the law. In addition to its methodological contributions, the paper addresses key challenges in forensic semiotics: subjectivity, reproducibility, and transparency. It shows how computational approaches can mitigate the "subjectivity effect" by offering replicable, data-driven analyses of linguistic behaviour. Courtroom admissibility is also supported by producing clear, step-by-step analytical records and exportable logs that document each analytic operation. Our contribution aligns with the aims of computational forensic linguistics to enhance the interpretive rigour and evidentiary validity of linguistic analyses.

Keywords: Authorship Analysis, Courtroom Admissibility, KWIC Concordancing, Linguistic Evidence, Shared N-gram Profiling, Similarity Assessment

1 Introduction

1.1 Context and Motivation

Forensic linguistics is the application of linguistic principles and methodologies to legal and investigative contexts. It encompasses a wide range of analyses, including accent identification [1], authorship attribution [2], deception detection [3], and sociolinguistic profiling [4]. One of its core applications is authorship analysis, where linguistic features such as syntactical, lexical, and stylistic choices are examined to determine the likely author of a disputed or anonymous text. In forensic investigations, authorship attribution plays a key role in cases involving *inter alia* threats [5], plagiarism [6], and cybercrime [7]. However, forensic linguistic investigations present multiple challenges, including data scarcity [8, 9], variability in writing style [5, 10, 11], and intentional obfuscation [12, 13].

In forensic authorship analysis, texts are often categorised as questioned or known texts. A questioned text (Q) is a document of uncertain or disputed authorship, such as an anonymous review, an online post, or a threatening note. A known text (K), on the other hand, is a document whose authorship is confirmed, typically written by a suspect or a person of interest. The primary objective of forensic authorship analysis is to determine shared distinctive linguistic similarities between Q and K, thereby establishing or refuting authorship.

Authorship investigations involve both linguistic and computational analyses of similarities and differences of Q and K texts to assess authorship likelihood. Linguistic analyses of digital texts focus on idiosyncratic variations in syntax, lexical choices and non-standard language. These analyses require a high degree of expertise and take a substantial amount of time. Computational analyses, such as n-gram analysis, frequency profiling, and stylometric techniques, can further enhance the comparison. Cross-textual similarity measures, such as cosine similarity, Jaccard similarity, and distance-based metrics, may provide quantitative assessments of similarity in computational analyses. Both lay and expert users can conduct stylometric analyses although expertise is needed to interpret the results.

1.2 Problem Statement

Existing concordancing tools widely used in corpus linguistics and forensic linguistic research suffer from several usability limitations that hinder efficient forensic analysis. One of the most significant issues is the inability to display Key-Word-In-Context (KWIC) results alongside the full-text view. In almost all cases, KWIC concordance lines are presented in a separate tab or window, necessitating constant switching between views. This fragmentation disrupts workflow and impairs the ability to analyse

linguistic patterns in their broader textual context, which is particularly problematic for forensic linguistic investigations where contextual meaning is paramount.

Another important usability issue is the absence of built-in part-of-speech (POS) tagging functionality. Many concordancing tools require users to preprocess their corpora with external POS taggers before conducting POS-related searches. This additional step increases the complexity of the analysis process, reducing accessibility for linguists with limited technical expertise.

Further complicating usability is the rigid structure of search constraints. Current concordancers often separate search functionalities for token-based, regex-based, and POS-based queries, requiring users to manually pre-select the appropriate search mode or preprocess input data before conducting searches. This separation restricts flexibility, as users must adjust their query strategy based on predefined system constraints rather than dynamically combining different search methods within the same submission field. The inability to fluidly integrate different types of searches forces users to modify their queries multiple times.

Beyond these core usability concerns, additional limitations affect the practicality of existing concordancers. Many tools require local installation, reducing accessibility for users who prefer web-based solutions. Additionally, the absence of a search history feature prevents users from easily refining previous queries, necessitating manual re-entry and slowing down iterative analysis. The overall lack of a user-friendly interface and efficient search refinement mechanisms limits the adaptability of these tools, making them less suitable for forensic applications. A significant shortcoming of existing systems is their inability to support shared n-gram comparison across questioned and known corpora. They also lack the functionality for conducting shared n-gram consistency checks within documents produced by the same author.

1.3 Aim and Scope

This paper introduces a bespoke web application for authorship analysis that incorporates: (1) a KWIC search system to extract occurrences of specific keywords from uploaded documents along with their surrounding context, and (2) an n-gram similarity comparison functionality whose results are displayed as a frequency table. This web app bridges accessibility and rigour by offering an interface simple enough for legal practitioners to use while maintaining the methodological precision required for scholarly scrutiny and courtroom critique.

1.4 Contributions

This paper makes four contributions to the extant research literature. First, it addresses the long-standing divide between methodologically robust but technically demanding bespoke authorship analysis programs and the simpler Graphical User Interface (GUI)-based concordancers used by non-specialists. We present a web-based system that allows legal practitioners and other non-technical users to conduct rigorous analyses through an intuitive interface, showing that accessibility and forensic precision need not be mutually exclusive. Second, whereas existing practice typically separates quantitative similarity measures from qualitative contextual interpretation,

we integrate shared n-gram profiling with KWIC-based inspection in a single application, reducing the analytical handover between tools and uniting computational detection with linguistic interpretation. This integration effectively streamlines the analytical workflow, minimising opportunities for human error arising from manual data transfer, data loss, or parameter inconsistencies across tools and file formats. Third, recognising that forensic linguistic analysis is often criticised for opacity and limited reproducibility, the system records the complete analytical history, allowing actions to be replicated, and enables export of results for transparent review and evidential disclosure. Fourth, although primarily designed for authorship attribution, this web app can be utilized to aid forensic linguistics in three canonical domains of forensic linguistics, namely: in language as evidence through similarity assessments between questioned and known documents, in legal interaction through the identification of recurrent questioning frames, and the language of the law through clause comparison and statutory drift analysis.

2 Background and Related Work

2.1 Concordancers and KWIC Tools in Corpus Linguistics

The concept of KWIC originated in the 1950s with the work of Hans Peter Luhn [14]. KWIC searches are highly informative in forensic authorship analysis, as they allow investigators to examine distinctive language patterns across different texts. KWIC concordancing remains one of the foundational techniques in corpus linguistics, providing a structured means of examining lexical patterns through the systematic alignment of a keyword and its immediate co-text [15, 16]. Originating in mid-twentieth-century information retrieval and later formalised in corpus linguistics, the KWIC display is valued for its ability to foreground recurrent lexical, collocational, and phraseological behaviour while allowing analysts to infer functional and pragmatic nuance from surrounding context. Concordancers have been used extensively in domains such as lexicography, discourse analysis, language pedagogy, phraseology, and register analysis, where the chief concern is interpretive insight rather than evidential accountability.

Over time, concordancer design has evolved to incorporate more sophisticated linguistic processing. Some systems now offer syntactic parsing and automatic multiword expression extraction [16, 17], while others support parallel concordancing for translation studies [18] or integrate morphological analyzers for languages with complex inflectional systems [19]. Despite these developments, concordancers largely retain their original epistemic model: they support exploratory, researcher-driven inspection of corpus evidence, favouring interpretive flexibility over procedural constraint.

However, the strengths of traditional concordancers expose limitations when they are repurposed for forensic linguistic work. First, standard KWIC systems treat each corpus as an isolated unit and therefore do not natively support cross-corpus comparison, an essential requirement when assessing similarity between questioned and known texts. Second, concordance lines are typically displayed apart from the full text, requiring analysts to switch between views, thereby slowing inspection and fragmenting interpretive reasoning. Third, very few concordancers preserve a traceable record of query parameters, search order, or output history, which means that results cannot

be reproduced with evidential certainty. Finally, because concordancers were designed for scholarly rather than legal purposes, no consideration was given for audit-ready record-keeping required for admissibility under Daubert- or Frye-style standards [20]. These limitations do not undermine their value in linguistic research, but they do constrain their suitability as standalone tools in forensic authorship analysis.

2.2 Stylometric and Statistical Authorship Tools

Parallel to concordancing, stylometry has developed as a computational approach to authorship attribution, particularly within literary studies. Stylometric methods typically rely on the extraction of statistical features—such as character or word n-grams, function-word frequencies, sentence length distributions, or POS tag sequences, which are then fed into classification algorithms or distance-based similarity metrics. A widely used framework in authorship analysis is the `stylo` [21] package in R, which provides pipelines for feature extraction, clustering, supervised classification, and visualisation. Similarly, `idiolect` [22] focuses on identifying distinctive lexical or stylistic features across multiple texts attributed to the same author. These tools support a high degree of methodological configurability and are well suited to experimental exploration in research settings.

Yet, for all their analytical power, stylometric environments pose three obstacles in forensic contexts. First, they presuppose familiarity with command-line workflows or scripting in R, which remains a significant barrier for legal practitioners, investigators, and expert witnesses who lack programming experience. Admittedly, since Large Language Models (LLMs) are capable of generating code, analysts may be tempted to delegate this task to them. However, programs produced by LLMs are not error-free, and without sufficient programming literacy to read and evaluate the code, the reliability of their output remains uncertain. Second, stylometric methods do not provide integrated mechanisms for contextual validation: stylometric distances may suggest similarity, but they do not reveal how a feature functions within its discursive environment, leaving interpretation to a separate concordancing step. Third, although `stylo` [21], `idiolect` [22], and related tools can output numerical measures and dendrograms, they lack embedded support for evidential traceability: parameters may be altered mid-workflow without record, rendering exact replication difficult and courtroom scrutiny problematic. As such, stylometric tools provide methodological rigour but lack the procedural scaffolding required for evidential admissibility, since understanding how input data are processed into output necessitates direct examination of the underlying program code.

2.3 Forensic-Specific Authorship Attribution Software

In response to the practical constraints of general corpus and stylometric tools, several purpose-built authorship attribution systems have been developed. Signature [23], for instance, aimed to assist forensic experts by providing frequency profiles of discriminating features, though its functionality remained limited to a narrow set of lexical statistics. Although Version 1.0 of Signature was released in 2004, there have been no updates since then.

NeoNeuro[1], a lightweight application, identifies shared n-grams across questioned and known texts and outputs percentage-based similarity scores. Its algorithms are opaque, rendering it unusable in forensic settings. The software, although still available for purchase, appears outdated with an interface design reminiscent of those from the previous millennium. While simple to use, the absence of contextual inspection restricts its utility in evidential reasoning, where analysts must justify not only that two texts share features, but also how and why those features matter pragmatically or stylistically.

The Java Graphical Authorship Attribution Program (JGAAP)[2] [24, 25] represents a more ambitious attempt to provide a menu-driven interface for multiple feature types and classification algorithms. JGAAP lowers the scripting barrier somewhat, but it still requires installation and configuration through GitHub. More importantly, the project is no longer actively maintained, raising concerns about software longevity, version-control integrity, and legal defensibility. In expert testimony, reliance on discontinued or opaque software can be grounds for evidential challenge, thereby ruling out this program for use in forensic contexts.

By contrast, the R packages `stylo` [21] and `idolect` [22] offer greater analytical flexibility but at the cost of usability. They presuppose competence in R, lack GUI-based logging, and require manual file handling, all of which increase cognitive overhead during casework. The widest gap among these tools is the absence of a combined confirmatory–exploratory workflow: some systems provide numerical similarity without context, while others provide context without similarity metrics. None of these applications integrates a cross-dataset frequency table of shared n-grams with a KWIC-based interpretive pathway, and none foreground reproducibility or auditability as primary design objectives.

2.4 Repurposed Corpus Tools in Forensic Practice

In practice, many forensic linguists continue to rely on general-purpose corpus tools such as AntConc [26], WordSmith Tools [27], LancsBox [28], and Sketch Engine [29]. These systems are powerful and well established, but none were designed with forensic evidential workflows in mind. AntConc, for example, offers KWIC searching and keyword analysis but lacks built-in cross-corpus comparison, POS-sensitive similarity inspection, or parameter-locked reporting. Its search history is not preserved, and datasets must be manually reloaded, disrupting continuity and precluding a verifiable chain of operations.

WordSmith Tools provides extensive statistical output but requires local installation and licences, limiting accessibility and complicating collaborative review. LancsBox supports POS-tagged searches and collocation networks but does not integrate frequency comparison across corpora, nor does it maintain an auditable record suitable for use in expert witness testimony. Sketch Engine is a subscriptiona-based service that is extremely capable at scale as it is optimised for billions of tokens. However, most forensic casework involves texts comprising hundreds or thousands of words rather than billions. Across all of these tools, the same constraints recur: fragmented

[1] https://neoneuro.com/products/authorship-attribution
[2] https://github.com/evllabs/JGAAP

workflow, absence of parameter locking, lack of evidential traceability, and limited suitability for non-technical users such as lawyers, police analysts, or judges.

As a result, forensic practitioners frequently resort to ad hoc multi-tool workflows. A typical process might involve running KWIC searches in AntConc, exporting frequency tables to Excel, computing similarity scores in SPSS, and then reconstructing an interpretive narrative in a word processor. Apart from being time-consuming, such workflows are error-prone and almost impossible to replicate precisely, particularly when intermediate steps are undocumented or overwritten [30]. From a methodological standpoint, this lack of transparency constitutes a form of evidential entropy: each unlogged step weakens the defensibility of the final conclusion.

2.5 Methodological Requirements in Forensic Linguistics

The limitations identified above are not merely inconveniences; they run counter to the methodological requirements of forensic practice. Unlike exploratory corpus linguistics, forensic analysis must be reproducible, reviewable, and reportable. Expert evidence must survive adversarial scrutiny, which requires a traceable record of all parameters, datasets, and procedural steps. Forensic linguists must therefore balance three interconnected obligations: analytical sensitivity, interpretive transparency, and procedural rigour. Analytical sensitivity requires the ability to detect subtle but meaningful similarities and differences between linguistic samples, even when they are concealed within stylistic variation or deliberate disguise. Interpretive transparency demands that such findings be explained in a way that makes the underlying linguistic mechanisms clear and accessible, showing precisely how identified patterns operate within their textual and situational contexts. Procedural rigour, in turn, ensures that every analytical step from data preparation and parameter selection to the generation and interpretation of results can be replicated, reviewed, and disclosed for evidential scrutiny. These three obligations form the ethical and methodological foundation of forensic linguistic practice, supporting both the credibility of the analyst and the admissibility of the evidence in judicial settings.

Two further constraints shape tool design in this domain. First, many forensic stakeholders are non-technical. Police analysts, lawyers, or court officials may need to inspect results without writing code or executing scripts. Second, interpretation in forensic linguistics is never purely computational: numerical similarity alone does not establish authorship. Contextual inspection is required to distinguish genuine stylistic alignment from coincidental overlap. Traditional corpus approaches often prioritise frequency as the principal indicator of significance, yet in forensic contexts salience [31], such as the prominence or distinctiveness of a linguistic feature within its communicative or situational context may be more meaningful [32]. A rare but striking collocation can carry greater evidential weight than a frequent but generic one, underscoring the need to balance quantitative frequency with qualitative interpretive salience when assessing linguistic similarity. Any viable all-in-one forensic tool must therefore integrate confirmatory measures such as shared n-gram statistics with exploratory interpretation such as KWIC context inspection.

2.6 Gap in Existing Approaches

Across the four strands reviewed—concordancers, stylometric and statistical tools, forensic attribution programs, and repurposed corpus software—no existing system combines cross-dataset similarity measurement, contextual validation, GUI-based usability, and audit-ready reproducibility. Concordancers provide context but not comparison; stylometric tools provide comparison but not context. Forensic-specific tools simplify interfaces but sacrifice analytical depth; repurposed corpus tools offer power but not evidential traceability. The absence of an integrated environment that unifies these functions for small, sensitive forensic datasets represents both a clear research gap and a niche for a forensic-first corpus tool. This is the gap that the present web application is designed to fill, providing a unified platform for contextual exploration and comparative analysis within a transparent and reproducible workflow.

3 System Description

3.1 Design Principles

The system must support a range of user interactions, including intuitive input mechanisms, comprehensive search functionalities, and interactive display of KWIC results. Users should be able to input textual data in various formats and conduct searches based on multiple linguistic parameters such as token matching, POS tagging, and regular expressions. The system should also provide robust result presentation and interpretation features, allowing users to save search outputs, apply custom labels, and export findings in various formats (e.g., `CSV`, `JSON`, `PDF`) for further analysis or record keeping. Additionally, the ability to refine and iterate searches within previously obtained results will enhance forensic investigations.

Performance requirements prioritise accuracy over speed, given that forensic linguistic analyses demand precise results rather than high throughput. While efficiency remains a consideration, scalability is typically not a primary concern in forensic linguistic applications due to the common challenge of data scarcity rather than large-scale data processing. Usability and accessibility are vital though, as many investigators and forensic linguists rely on graphical user interfaces (GUIs) rather than command-line interfaces (CLIs). The system must be designed to accommodate users with varying levels of technical expertise, ensuring an intuitive and efficient workflow. Data privacy and security are also paramount, given the sensitivity of forensic investigations.

To effectively support forensic investigations, the system must offer a high degree of customisability, allowing users to tailor search parameters and analytical functions to specific case requirements. The system should also incorporate traceability and logging features to maintain a comprehensive record of all search activities, ensuring reproducibility and auditability in forensic analysis.

3.2 System Architecture

Fig. 1 provides a visual illustration of the system architecture, which comprises a frontend for the web application user interface and a backend for search and data

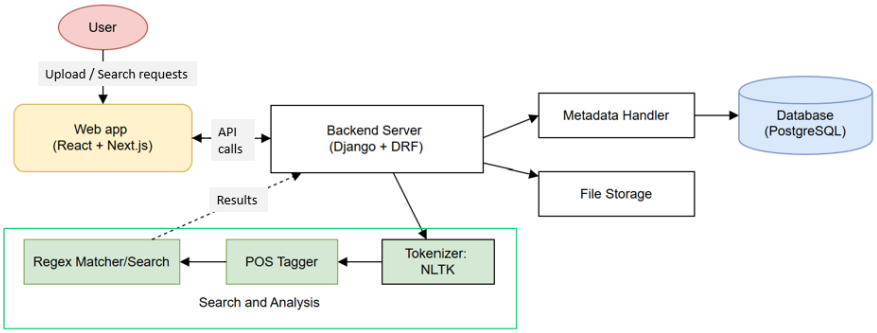


Fig. 1 System architecture of the authorship analysis web app (created by the authors).

management. When a user uploads a file or corpus, the backend registers its metadata (e.g., file name, author name, description) in a database, while storing the actual file in server-side storage. In response to a search request, the system performs real-time analysis on the target document, including tokenisation and POS tagging, and matches it against the user-specified search conditions.

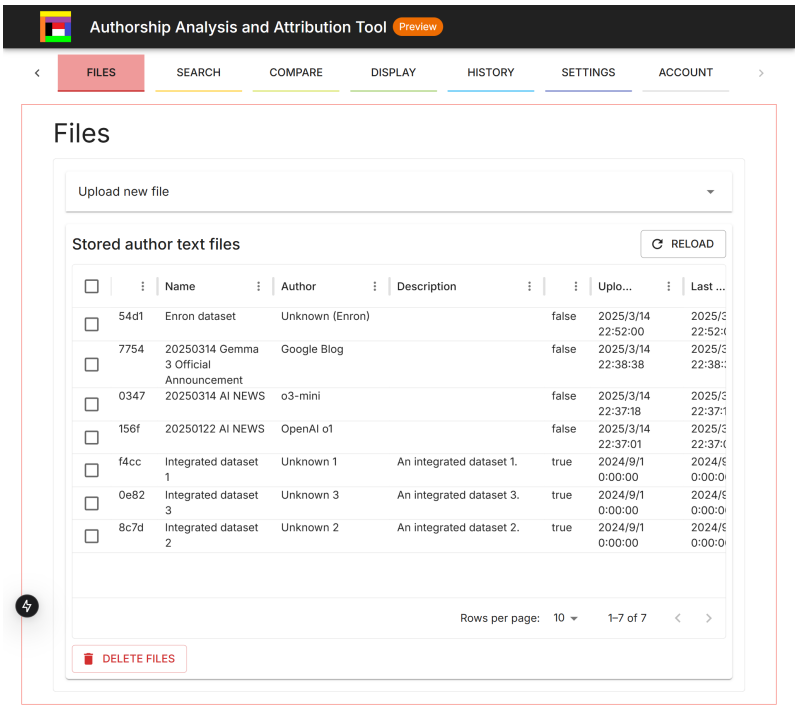


Fig. 2 Screenshot of the file management tab.

The GUI offers an intuitive interface for managing documents, conducting KWIC searches, and comparing n-grams between and among corpora. Users can upload, list, or delete files, and specify search conditions as needed. Fig. 2 shows the file management tab in which users can create, rename, update and delete files. The search results highlight the keywords within their immediate context window, providing a clear visualisation of how terms are used across the text. The main features of the system are the KWIC search and the shared n-gram comparison module. The former enables contextual exploration of lexical and structural patterns within individual texts, while the latter automatically identifies and visualises recurring bigrams and trigrams across uploaded documents, revealing patterns of overlap and divergence.

3.3 Core Functionalities

3.3.1 KWIC Searches

A KWIC search can identify occurrences of a specified keyword in multiple datasets along with its surrounding words, providing insight into its usage and contextual meaning within and between datasets. KWIC search results show concordance lines in a way that maintains the keyword (also known as node or hit) centrally positioned within a fixed-width window, allowing for easy comparison of linguistic patterns. By aligning keywords with their surrounding contexts, analysts can detect idiosyncratic tendencies. KWIC searches can reveal unusual stylistic markers, such as consistent misspellings or preferred collocations.

Our system is similar to existing KWIC approaches in that it can search sequences of multiple tokens; however, it also supports wildcard specifications at the token level and commonly-used regular expressions, thus providing more flexibility. For instance, users can apply a regular expression such as `th.*` to a portion of a token and combine it with POS tags to refine the search. This capability is distinctive in that it integrates token-level matching with both regular expression patterning and POS tag searches within a single query, thereby supporting more flexible and fine-grained analyses of grammatical structures. Specifically, queries like `th.* JJ NN` can extract sequences such as *The perfect timing*, capturing any token beginning with *th* followed by an adjective and a noun.

Fig. 3 shows an example of a search incorporating word tokens, a part of speech and a regular expression. By default the first 20 instances are displayed with an option to load more. The keyword is colourised and aligned vertically. The raw text of the dataset is displayed underneath the results, providing users with the opportunity to explore the context and conduct other searches within the raw data while still viewing the KWIC results.

This version currently employs standard Python libraries and the Natural Language Toolkit (NLTK) [33] to tokenise documents and assign Penn Treebank POS tags before performing regular expression and POS-based pattern matching. As the system does not build an index in advance, the search time depends on the size of the documents.

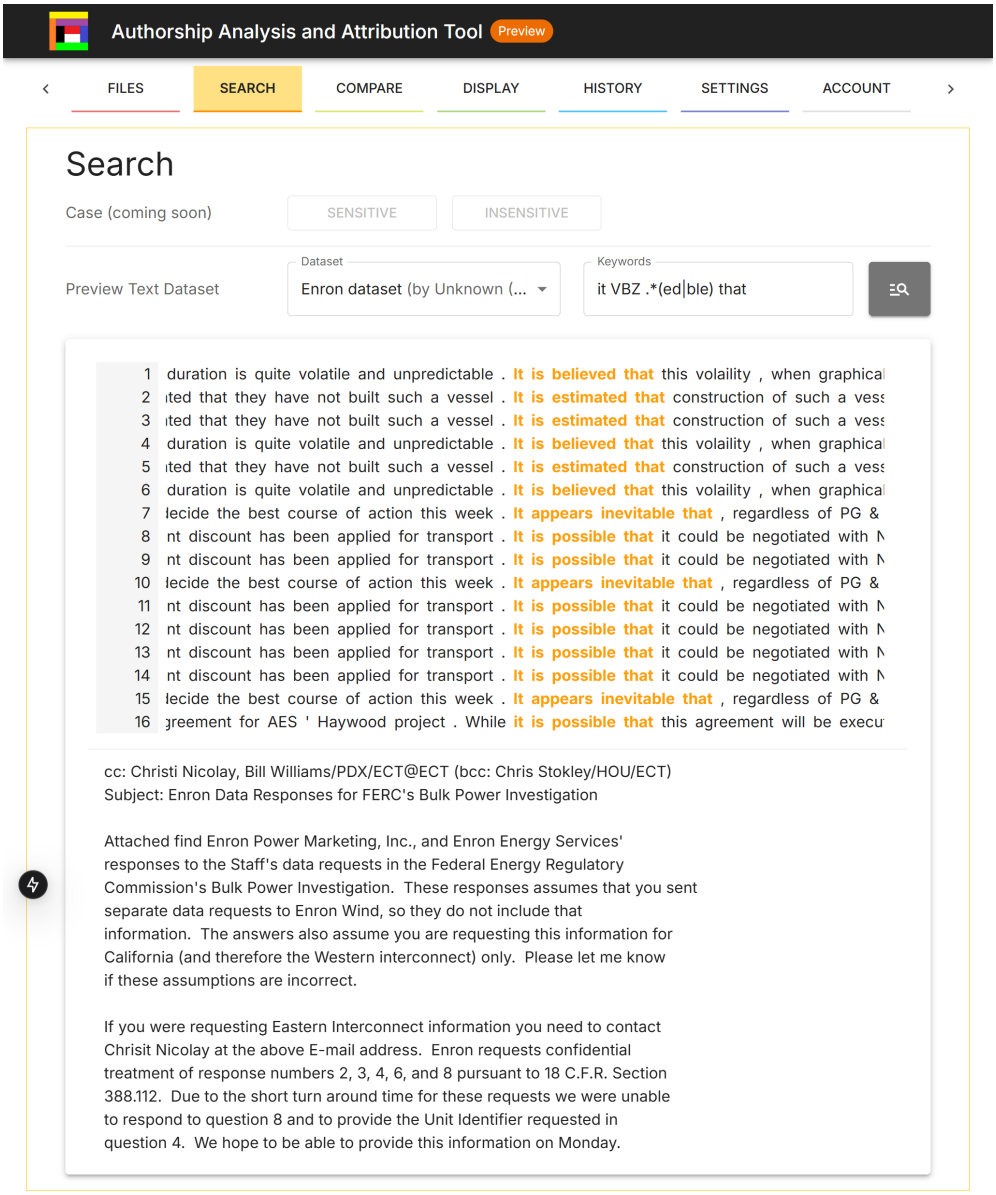


Fig. 3 Screenshot of the KWIC results in the Search tab with full-text view displayed underneath the concordance lines. The node is identified through a search comprising word tokens, POS tag and a regular expression.

3.3.2 Shared N-gram Frequency Table

The system automatically counts recurring sequences of words, known as n-grams, with bigrams ($n = 2$) and trigrams ($n = 3$) used by default. These counts are displayed in a frequency table where each row represents an n-gram and each column represents a dataset. The cells show how many times each n-gram occurs in each dataset. When the same n-gram appears in more than one dataset, its corresponding cells are highlighted in yellow, allowing users to quickly identify shared expressions across texts.

The system incorporates two complementary similarity-detection functions within its n-gram analysis module: the consistency checker (intra-corpus comparison) and the cross-corpus comparison (inter-corpus comparison) [34]. Both rely on the same n-gram engine, which extracts unigrams, bigrams, and trigrams from each uploaded text and calculates their occurrence frequencies and distribution across documents.

In the consistency checker mode, the system examines linguistic stability within a single author's corpus. It identifies recurring n-grams that appear in multiple files, ranks them by the number of documents in which they occur, and presents them in descending order of recurrence. This allows analysts to observe potentially distinguishing linguistic patterns that remain consistent across an author's writing.

In the cross-corpus comparison mode, one document or corpus is designated as the Questioned (Q) text, while additional corpora are designated as Known (K). The system compares n-gram profiles between Q and each K corpus, highlighting overlapping sequences. Shared n-grams are ranked by their frequency within the questioned text, with corresponding frequencies in the known corpora visually emphasised to facilitate rapid comparison.

These two modes provide complementary insights: intra-corpus analysis reveals stylistic consistency within a single author's production, whereas inter-corpus comparison identifies linguistic overlap that may indicate shared authorship or stylistic influence.

4 Methodological Rationale

The design of the system is not merely a matter of software engineering but of methodological alignment: any tool intended for forensic authorship analysis must satisfy both the theoretical expectations of the discipline and the practical realities of casework. Forensic linguistics is positioned between academic inquiry and evidential procedure, meaning that analytical methods must be simultaneously interpretable, defensible, and operationally usable. This section, therefore, clarifies the rationale behind the system architecture, showing how core design choices are grounded in established principles of forensic method. Our rationale rests on five foundational concepts. First, the choice of approach is justified in light of the enduring tension between rigour and accessibility. Second, the workflow is framed as an intentional pairing of confirmatory and exploratory techniques. Third, transparency and reproducibility are embedded as design constraints rather than optional features. Fourth, the discussion addresses the importance of evidential restraint. Finally, although the system is developed with a primary focus on authorship analysis, it is designed with applicability across the

broader domains of forensic linguistics in mind. This broader orientation provides a natural segue into the case studies that follow.

The system is designed to bridge two essential yet competing priorities: methodological rigour and operational accessibility, namely: ease of use via a clear, code-free graphical interface. Fully fledged stylometric pipelines, while powerful, require substantial statistical and technical expertise, whereas traditional GUI-based concordancers offer usability at the expense of forensic precision and auditability. The present system adopts a middle path by selecting techniques that are both tractable for non-specialists and sufficiently systematic to withstand scholarly scrutiny and courtroom critique.

4.1 Confirmatory–Exploratory Workflow

The workflow rests on a deliberate pairing of two complementary procedures. The shared n-gram frequency table offers a confirmatory signal of lexical overlap across datasets, providing a transparent, defensible entry point into similarity assessment. KWIC inspection then supplies the exploratory nuance required to distinguish systematic stylistic correspondence from coincidental similarity. The sequence is intentional: counts first, contexts second, interpretation last. By beginning with systematically identified shared n-grams, the workflow anchors interpretation in quantifiable evidence, thereby constraining subjective bias.

4.2 Transparency, Replicability, and Auditability

Because forensic linguistics is frequently challenged for opacity, the system embeds traceability at the level of process rather than merely output. All analyses are recorded in an auditable history log. Dataset hashes, timestamps, and provenance metadata ensure that the complete analytical history can be reconstructed, reviewed, and disclosed. This design directly targets the "subjectivity effect" and aligns the workflow with expectations of expert evidence under Daubert- or Frye-style standards [20].

4.3 Evidential Restraint

Patterns alone do not constitute proof; they are evidential indicators, which require calibration and caution. The shared n-gram analysis is, therefore, framed as indicative rather than determinative. Shared n-grams indicate stylistic convergence, though such overlap may arise as much from genre and content similarity as from authorial style [35, 36]. KWIC inspection of the concordance lines and, where necessary, the wider co-text can help guard against false positives by reintroducing context. The system is built not to claim certainty, but to provide objective evidence that the analyst can interpret.

4.4 Domain-General Applicability

Although the core analytical procedure, namely combining systematic n-gram detection with contextual validation, remains constant, its utility extends across the three canonical domains of forensic linguistics [37]. In language as evidence, our web app assists in assessing authorship consistency between questioned and known texts; in legal interaction, it can help reveal recurring institutional and individual patterns of

language use; and in the language of the law, the workflow supports, for example, comparison and detection of unmarked textual drift between or among different drafts.

5 Evaluation and Case Studies

This section presents three domain-specific use cases that demonstrate how the web application supports the three core domains of forensic linguistic practice: (1) language as evidence, (2) interaction in legal settings, and (3) language of the law. These three domains represent the semiotic breadth of forensic linguistics from the drafting of legal texts, through the dynamics of institutional interaction, to the treatment of language in investigative and judicial contexts. As these domains differ in textual form, communicative function, and evidential purpose, they serve as a triadic testbed for assessing the versatility of the web application. Each use case follows a common structure consisting of an objective, dataset, procedure, key findings, and forensic relevance.

5.1 Use Case 1. Language as Evidence: Authorship Attribution and Similarity Assessment

Objective

This use case demonstrates how the web application can be used to assess authorship similarity in a scenario where a questioned email (Q) must be evaluated against two candidate authors, Author A and Author B. The aim is to determine whether patterns of lexical overlap and contextual usage provide evidence of stylistic alignment.

Dataset

To avoid confidentiality concerns, all texts in this section are synthetic; however, the workflow is directly transferable to real case material. The dataset comprises two short professional email memos written by Author A (A1, A2), two by Author B (B1, B2), and one questioned email (Q) whose authorship is unknown. The five texts exhibit distinct but realistic register profiles: Author A writes concise, bureaucratic memos characterised by formulaic business trigrams, whereas Author B adopts a hedged academic–legal register marked by longer, clause-heavy expressions. The questioned memo appears, at first glance, more stylistically aligned with Author A, but it contains several embedded trigrams that are characteristic of Author B, creating a plausible case of stylistic leakage. The full texts are reproduced below.

Sample Texts Used in the Analysis

Q (Questioned Memo)
Please be advised that the attached access schedule is effective immediately and remains non-exclusive and non-transferable. In light of the fact that prior approvals have lapsed, it is reasonable to assume that updated forms will be required. At this particular juncture, compliance should proceed in accordance with existing guidance until new instructions are issued.

A1 (Known Author A)
Please be advised that all software access requests must be submitted in writing. The license is non-exclusive and non-transferable, and prior written approval is required for any deviation. This procedure is effective immediately and applies to all internal departments without exception.

A2 (Known Author A)
Please note that renewal forms must be completed before the end of the current quarter. Access remains non-exclusive, non-transferable, and subject to prior written approval. All changes take effect immediately unless otherwise stated in company policy.

B1 (Known Author B)
In light of the fact that the revised protocol has not yet been circulated, it is reasonable to assume there will be procedural uncertainty. At this particular juncture, one might infer that interim guidance is required, and it should be emphasised that compliance depends on departmental initiative.

B2 (Known Author B)
At this particular juncture, it is reasonable to assume that the lack of formal notification will create workflow ambiguity. In light of the fact that no follow-up memo has been issued, one might infer that clarification is forthcoming.

Procedure

All five documents were uploaded to the web application and processed using the shared trigram comparison function. The resulting frequency table was generated and sorted according to trigram overlap between the questioned text and each candidate author. Following this, targeted KWIC searches were conducted on the shared trigrams to determine whether the observed overlaps reflected generic formulaic language or systematic stylistic patterning within each author's writing.

Key Findings

The shared trigram table (Table 1) reveals a dual pattern of alignment. The questioned email contains a small number of trigrams associated with Author A, primarily formulaic business expressions such as *please be advised* and *effectively immediately*. However, a substantially larger and more internally consistent set of trigrams aligns

with Author B, including extended clause-building sequences such as *in light of the fact*, *it is reasonable to assume*, and *at this particular juncture*.

While Author A shares four trigram matches with the questioned text, Author B exhibits a much higher total frequency of overlap (26 instances), reflecting repeated use of structurally similar expressions across both B1 and B2. This pattern suggests not only overlap with the questioned text but also internal stylistic consistency within Author B's writing.

Table 1 Shared trigrams between Q and K datasets.

Trigram	Questioned	Author A	Author B
`please be advised`	1	1	0
`be advised that`	1	1	0
`is effectively immediately`	1	1	0
`effectively immediately and`	1	1	0
`. in light`	1	0	1
`in light of`	1	0	2
`light of the`	1	0	2
`of the fact`	1	0	2
`the fact that`	1	0	2
`, it is`	1	0	2
`it is reasonable`	1	0	2
`is reasonable to`	1	0	2
`reasonable to assume`	1	0	2
`to assume that`	1	0	1
`. at this`	1	0	1
`at this particular`	1	0	2
`this particular juncture`	1	0	2
`particular juncture ,`	1	0	2
Total	**18**	**4**	**26**

To determine whether these overlaps are stylistically meaningful rather than coincidental, KWIC inspection was performed on the shared trigrams. The KWIC output (Fig. 4) shows that the expressions associated with Author B occur as part of extended, clause-building structures that recur across both B texts and the questioned memo. These patterns function as cohesive rhetorical units rather than isolated lexical coincidences, suggesting a habitual stylistic preference.

Forensic Relevance

This use case illustrates how the combined workflow supports evidential reasoning in authorship analysis. The shared n-gram profile provides a transparent quantitative signal of overlap, while the KWIC inspection supplies the contextual layer necessary to evaluate the functional role of those patterns. This enables the analyst to distinguish superficial lexical similarity from systematic stylistic alignment. In forensic casework, such integration is important, as it allows expert conclusions to be grounded in both measurable evidence and interpretable linguistic behaviour, thereby strengthening the defensibility of authorship inferences.

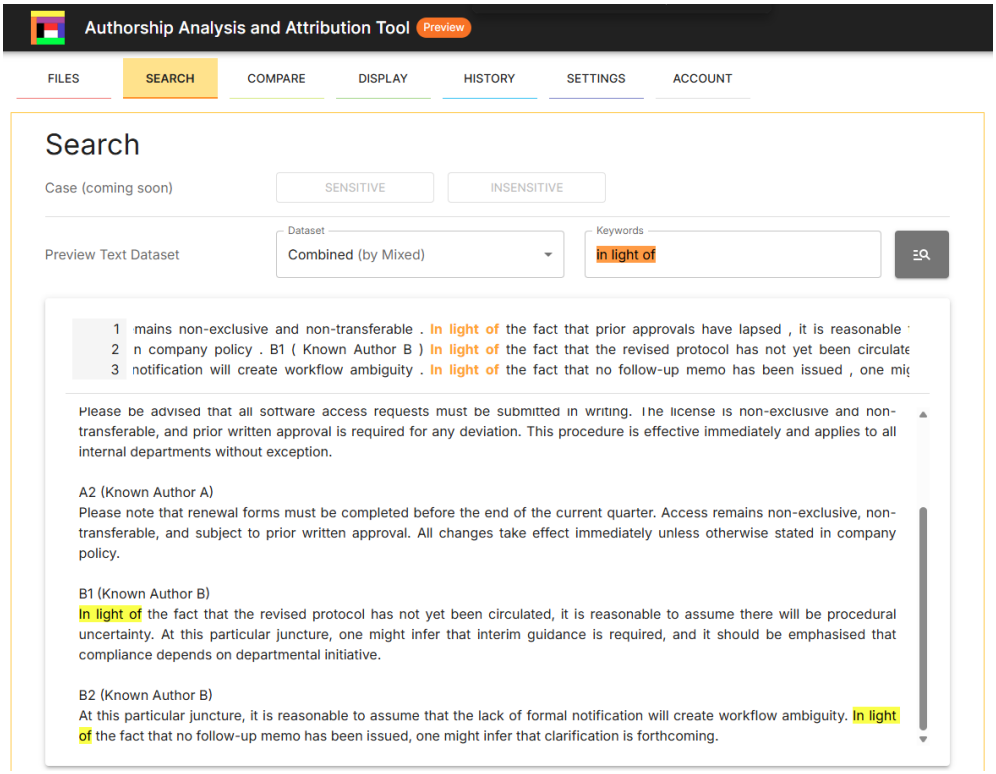


Fig. 4 Screenshot of the KWIC view showing a keyword search. Note the simultaneous display of concordance lines and full-text context.

5.2 Use Case 2. Interaction in Legal Settings: Reformulation and Confirmation in Police Interviews

Objective

This use case demonstrates how the web application can be applied to the analysis of institutional questioning in police interview discourse, with a focus on identifying reformulation and confirmation strategies used by interviewers to shape or constrain a suspect's narrative. The aim is to determine whether recurrent linguistic patterns can be systematically detected and interpreted as evidence of interactional steering.

Dataset

A single synthetic police interview transcript is used to model the phenomenon of the *reformulation–confirmation loop*. The dataset consists of alternating turns between an interviewer (INT) and a suspect (SUS), designed to reflect realistic features of institutional dialogue, including partial responses, hedging, and incremental clarification. The transcript was constructed to include repeated interviewer reformulations and

subsequent suspect confirmations, allowing the analytical workflow to be demonstrated under controlled conditions while preserving interactional plausibility.

INT: Could you tell me what happened after you left the station?
SUS: I walked straight to the bus stop, didn't talk to anyone.
INT: Right, so what you're saying is you went directly to the bus stop and avoided people?
SUS: I guess, yeah, I just waited there.
INT: And just to confirm, you didn't speak to anyone on the way?
SUS: No, not really.
INT: So what you're saying is you had no contact with anyone between the station and the bus stop?
SUS: I mean, someone asked me for a lighter, but I didn't have one.
INT: Okay, just to confirm, you're saying you did not engage in conversation with that person?
SUS: I said I didn't have a lighter, that's it.
INT: So you're saying it wasn't really a conversation, just a refusal?
SUS: Yeah, I guess.
INT: And just to confirm again, there was no further exchange?
SUS: No.
INT: So you're saying that you only spoke one sentence and then stopped talking?

Procedure

The transcript was uploaded to the web application as three datasets: interviewer corpus, suspect corpus and combined corpus. A KWIC search was performed for the keyword *saying* to identify recurring reformulation structures used by the interviewer. Following this, the shared n-gram comparison function was applied to compare trigram distributions between interviewer and suspect turns, allowing for quantitative assessment of overlap and divergence in phraseological patterns.

Key Findings

The KWIC output (Fig. 5) reveals a highly systematic use of reformulation phrases by the interviewer, particularly patterns such as *so you're saying* and *so what you're saying is*. These constructions repeatedly reframe the suspect's prior statements into more explicit or categorical propositions.

The trigram comparison further shows that there is minimal direct phraseological overlap between interviewer and suspect speech. Out of 62 trigrams identified in the dataset, only three are shared between the two speakers. This indicates that the interviewer's reformulations are not simple repetitions of the suspect's words, but are instead constructed paraphrases that are likely to reshape the propositional content of the interaction.

Inspection of the dialogue confirms that the interviewer progressively narrows the interpretive space of the suspect's responses. For example, the statement *I walked straight to the bus stop* is reformulated into a stronger claim of avoiding all contact, while a minimal exchange *someone asked me for a lighter* is subsequently reframed as *you did not engage in conversation*. Through successive reformulation and confirmation, the suspect is led to endorse increasingly restrictive interpretations of their own actions.

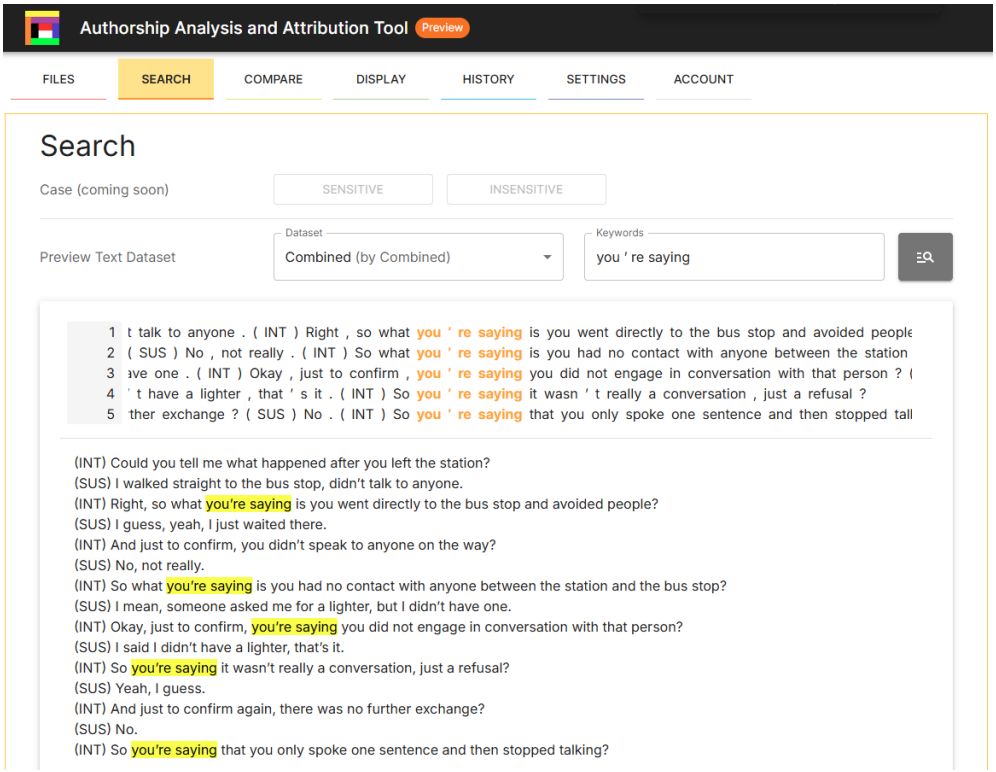


Fig. 5 Screenshot of the KWIC output for the keyword *saying*, showing repeated reformulation of the suspect's earlier answers.

Forensic Relevance

This use case demonstrates how the combined workflow can reveal interactional patterns that may not be immediately visible through unaided linear reading alone. The integration of KWIC analysis and n-gram comparison provides both qualitative and quantitative evidence of interviewer-led reformulation, allowing analysts to distinguish between a suspect's original wording and institutionally mediated paraphrases. In forensic practice, such analysis is directly relevant to the evaluation of interview fairness, the identification of suggestibility, and the assessment of evidential reliability in recorded statements. It also provides a transparent basis for expert testimony, particularly in cases where the linguistic construction of a narrative may influence judicial interpretation.

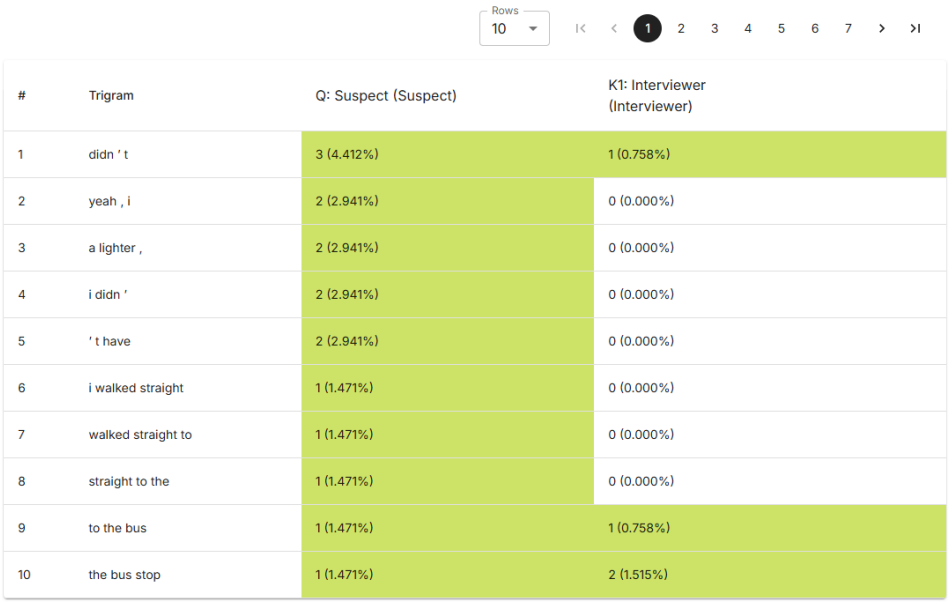


Fig. 6 Screenshot of the comparison output showing limited trigram overlap between interviewer and suspect turns.

5.3 Use Case 3. Language of the Law: Detecting Clause Drift in Software Licensing

Objective

This use case focuses on the domain of the language of the law, where even minor textual adjustments can carry disproportionately large legal effects. The objective is to determine whether the web application can detect *inter alia* wording drift, boilerplate reuse, and covert alteration in a standard software licence grant clause across multiple revision cycles.

Dataset

The dataset comprises seven versions (v.1–v.7) of a software licence clause created to simulate iterative drafting. Versions v.1 to v.4 preserve the same legal force but employ paraphrase, reordering, or lexical substitution, which is analogous to routine redrafting by different legal teams. Versions v.5 to v.7 introduce subtle but legally significant changes, including narrower usage scope, restrictions on sublicensing, and a transfer of liability to the licensee. All seven texts are shown below.

Dataset Texts (v.1–v.7)

v.1. The Licensee is hereby granted a non-exclusive, non-transferable licence to use the Software solely for its operational activities, subject to full compliance with the terms of this Agreement.

v.2. The Licensee is permitted to make use of the Software on a non-exclusive and non-transferable basis, provided that such use conforms to all obligations set out in this Agreement.

v.3. This Agreement grants the Licensee a non-exclusive, non-assignable right to utilise the Software for its business activities, contingent upon adherence to the contractual provisions herein.

v.4. The Licensor authorises the Licensee, on a non-exclusive and non-transferable basis, to operate the Software solely in connection with the Licensee's internal operations, subject to the terms below.

v.5. The Licensee is granted a non-exclusive, non-transferable licence to use the Software *for internal business purposes only*, and only in accordance with the conditions of this Agreement.

v.6. The Licensee may use the Software on a non-exclusive, non-transferable basis, but *shall not sublicense, distribute, or make the Software available to any third party* in any form.

v.7. The Licensee is granted a non-exclusive, non-transferable right to use the Software, *at the Licensee's sole responsibility and expense*, and subject to the full conditions set out herein.

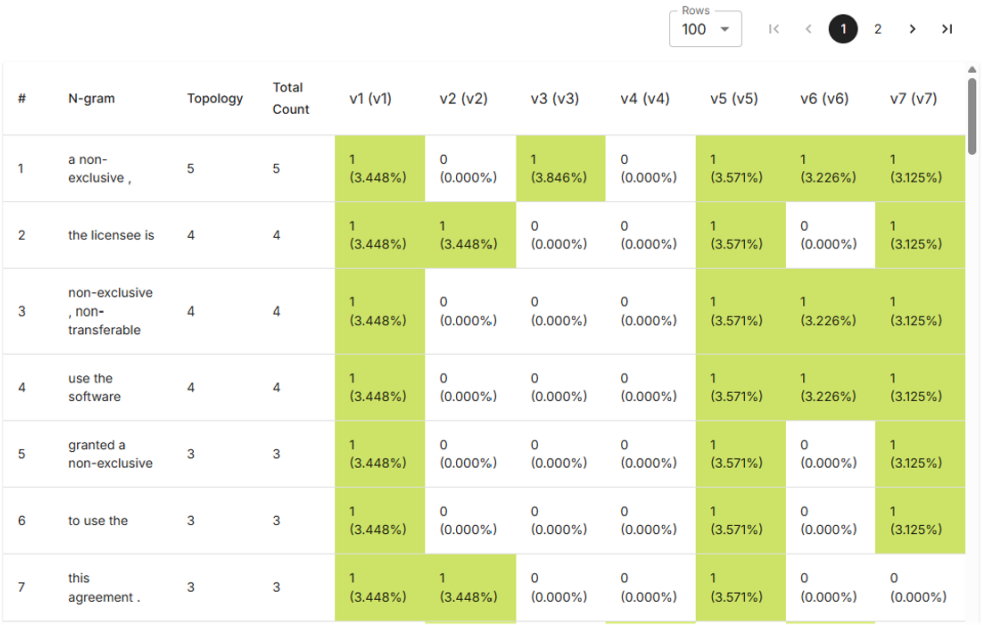
Note: Italicized text is used to highlight semantic shifts occurring in the later versions.

Procedure

All versions were uploaded individually to the web application and processed using the shared n-gram consistency tool. The resulting frequency table was sorted by n-gram frequency across the different versions. Following this, targeted KWIC searches were run on terms with recognised forensic salience (e.g., *use*, *non-exclusive*, *transfer*, *third party*, *sublicense*) to examine how these terms function within their immediate and wider co-text.

Key Findings

Figure 7 shows the output of the consistency function of the web application, which lists shared n-grams according to their frequency across datasets. The results show a rapid loss of lexical similarity between versions: only one trigram is shared across five versions, while a small number of additional trigrams are shared across four versions. Out of a total of 133 trigrams, only 26 occur in more than one version, indicating that even closely related drafts diverge substantially at the phraseological level.



#	N-gram	Topology	Total Count	v1 (v1)	v2 (v2)	v3 (v3)	v4 (v4)	v5 (v5)	v6 (v6)	v7 (v7)
1	a non-exclusive ,	5	5	1 (3.448%)	0 (0.000%)	1 (3.846%)	0 (0.000%)	1 (3.571%)	1 (3.226%)	1 (3.125%)
2	the licensee is	4	4	1 (3.448%)	1 (3.448%)	0 (0.000%)	0 (0.000%)	1 (3.571%)	0 (0.000%)	1 (3.125%)
3	non-exclusive , non-transferable	4	4	1 (3.448%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	1 (3.571%)	1 (3.226%)	1 (3.125%)
4	use the software	4	4	1 (3.448%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	1 (3.571%)	1 (3.226%)	1 (3.125%)
5	granted a non-exclusive	3	3	1 (3.448%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	1 (3.571%)	0 (0.000%)	1 (3.125%)
6	to use the	3	3	1 (3.448%)	0 (0.000%)	0 (0.000%)	0 (0.000%)	1 (3.571%)	0 (0.000%)	1 (3.125%)
7	this agreement .	3	3	1 (3.448%)	1 (3.448%)	0 (0.000%)	0 (0.000%)	1 (3.571%)	0 (0.000%)	0 (0.000%)

Fig. 7 Screenshot of the output of the consistency function showing shared and unshared trigrams across versions.

A similar consistency comparison for bigrams revealed that only three bigrams were present in all versions, namely *the licensee*, *the software*, and *a non-exclusive*. Versions v.1 and v.5–v.7 appear to share more similarities than versions v.2–v.4 according to bigram and trigram analysis. Analysis of the KWIC output for the node keyword *non-exclusive*, however, confirms that this condition permeates all versions.

The shared n-grams from v.5 onwards appear to signal a shift from stylistic redrafting to semantic alteration. Specifically, the emergence of phrases such as *for internal business purposes only*, *shall not sublicense*, and *sole responsibility* marks a transition from neutral paraphrase to legally consequential modification. KWIC inspection shows that v.5 restricts the licence to internal use, v.6 introduces an explicit prohibition against sublicensing or distribution, and v.7 transfers financial and legal responsibility to the licensee.

A KWIC search for the phrase *use the software* further reveals how the surrounding conditions change across versions, illustrating how wording adjustments introduce new legal restrictions and obligations.

Forensic Relevance

This use case demonstrates how the combined workflow allows the analyst to distinguish innocuous linguistic drift from materially altered meaning. The shared n-gram interface provides an immediate diagnostic signal of divergence, while the KWIC view supplies the interpretive context needed to reveal how terms are recontextualised

Fig. 8 Screenshot of the KWIC output showing variation in the use of the phrase *use the Software* across versions.

across versions. In forensic and legal practice, such analysis is directly relevant to contract review, drafting audits, and expert testimony, where the ability to identify covert clause modification can have significant legal consequences.

6 Discussion

The evaluation across three distinct forensic domains shows that our web app successfully reconciles accessibility with analytical rigour. The shared n-gram table provides a transparent entry point into similarity assessment, while the KWIC view restores the contextual layer needed for evidential interpretation. This pairing supports both confirmatory and exploratory reasoning within a single interface, reducing the need to switch between multiple standalone analytical tools, scripting environments, and data-processing platforms. The parameter tracking and exportable history logs address concerns about reproducibility and traceability in forensic work, ensuring that every step can be reconstructed for peer review, disclosure, or courtroom scrutiny. The use cases show that the workflow is not limited to authorship attribution, but can be applied to other domains of forensic linguistics.

The workflow remains descriptive rather than inferential and therefore does not replace methods that produce calibrated likelihood estimates or quantified error rates.

Shared n-gram overlap is a proxy for stylistic similarity, but it may also capture topical repetition, boilerplate formulae, or organisational templates, particularly in institutional registers. The default POS tagging and tokenisation pipeline is sufficient for English prose, but may require adaptation for multilingual or morphologically rich datasets. The synthetic datasets used in this paper are effective for exposition but do not fully replicate the noise, obfuscation, genre instability, and adversarial manipulation present in real case material; they were intentionally kept small to foreground the core concepts rather than to approximate fully realistic datasets. As with all corpus-derived evidence, the interpretive burden rests on the analyst, who must distinguish meaningful convergence from coincidental overlap and avoid overclaiming based on numerical or visual salience alone.

The system has practical value for expert witness testimony, investigative screening, and methodological training. In legal drafting audits, it enables rapid detection of covert clause alteration; in institutional discourse analysis, it exposes interviewer scripting and reformulation loops; and in authorship comparison, it provides both the numerical profile and the contextual justification expected in expert testimony. The exportable tables and KWIC captures can be inserted directly into reports, thereby reducing transcription effort and increasing evidential clarity. For teaching and professional development, the interface offers a scaffolded pathway from surface-level pattern recognition to context-sensitive interpretation, helping students, lawyers, and police analysts to understand how linguistic evidence is constructed. Because the workflow is transparent and replicable, it may also be used to demonstrate best practice in expert reasoning and disclosure.

Future development will focus on extending the feature set beyond token-based n-grams to include hybrid token–POS patterns, dispersion metrics, collocational profiles, and optional integration with different POS-taggers such as SpaCy. Validation studies will include known-error-rate estimation, inter-analyst agreement testing, and evaluation on adversarially modified datasets.

7 Conclusion

This paper presented a web-based authorship analysis system designed to meet the dual demands of forensic casework: accessibility for non-technical users and methodological rigour fit for evidential scrutiny. The three use cases demonstrated that the combined workflow of shared n-gram profiling and KWIC-supported contextual inspection is both flexible and domain-general. In legal drafting, it exposed covert clause modification; in police interviewing, it revealed patterned reformulation strategies; and in authorship comparison, it distinguished genuine stylistic alignment from superficial lexical overlap. These results show that the value of the system lies not merely in its intuitive interface, but in the way it structures forensic reasoning: counts first, context second, interpretation third.

The evaluation further illustrates that the tool addresses long-standing weaknesses in existing workflows, which typically require analysts to move between multiple applications, reconstruct evidential logic manually, and rely on unlogged intermediate steps. By embedding parameter locking, exportable logs, and a traceable run manifest, the

system operationalises the core principles of reproducibility, transparency, and disclosure. In doing so, it contributes not only a piece of software, but a replicable analytical model for how computational evidence can be derived, documented, and defended.

More broadly, the work reaffirms that forensic linguistics does not require a choice between interpretive nuance and computational structure. The system serves as a methodological scaffold by making good practice easier to perform, easier to explain, and easier to teach.

Future development will refine the technical implementation, but the central contribution of this paper is already complete: a demonstrated, domain-spanning, audit-ready approach that shows how accessible tools can support rigorous forensic analysis without compromising evidential standards.

References

[1] Mikhailava, V., Lesnichaia, M., Bogach, N., Lezhenin, I., Blake, J., Pyshkin, E.: Language accent detection with CNN using sparse data from a crowd-sourced speech archive. Mathematics **10**(16), 2913 (2022) https://doi.org/10.3390/math10162913

[2] Rocha, A., Scheirer, W.J., Forstall, C.W., Cavalcante, T., Theophilo, A., Shen, B., Carvalho, A.R., Stamatatos, E.: Authorship attribution for social media forensics. IEEE transactions on Information Forensics and Security **12**(1), 5–33 (2016) https://doi.org/10.1109/TIFS.2016.2603960

[3] Constâncio, A.S., Tsunoda, D.F., Silva, H.d.F.N., Silveira, J.M.d., Carvalho, D.R.: Deception detection with machine learning: A systematic review and statistical analysis. PLOS One **18**(2), 0281323 (2023) https://doi.org/10.1371/journal.pone.0281323

[4] Perkins, R.C.: The application of forensic linguistics in cybercrime investigations. Policing: A Journal of Policy and Practice **15**(1), 68–78 (2021) https://doi.org/10.1093/police/pay097

[5] Nini, A.: Register variation in malicious forensic texts. The International Journal of Speech, Language and the Law **24**(1), 99–126 (2017) https://doi.org/10.1558/ijsll.30173

[6] Coulthard, M., Johnson, A., Kredens, K., Woolls, D.: Plagiarism four forensic linguists' responses to suspected plagiarism. In: Coulthard, M., Johnson, A. (eds.) The Routledge Handbook of Forensic Linguistics, pp. 551–566. Routledge, London (2010)

[7] Blake, J.: Online crime in the metaverse: A study on classification, prediction, and mitigation strategies. In: Elshenraki, H. (ed.) Forecasting Cyber Crimes in the Age of the Metaverse, pp. 66–77. IGI Global Scientific Publishing, Hershey, PA (2024). https://doi.org/10.4018/979-8-3693-0220-0.ch004

[8] Wright, D.: Corpus approaches to forensic linguistics: Applying corpus data and techniques in forensic contexts. In: The Routledge Handbook of Forensic Linguistics, pp. 611–627. Routledge, London (2020)

[9] Wright, D.: Corpus Approaches to Discourse in Forensic and Legal Contexts. Routledge, London (2025)

[10] Olsson, J., Luchjenbroers, J.: Forensic Linguistics. A&C Black, London (2013)

[11] Nini, A.: Codal variation theory as a forensic tool. In: Bridging the Gap(s) Between Language and the Law: Proceedings of 3rd European Conference of the International Association of Forensic Linguistics, pp. 31–41 (2013). Faculdade de Letras da Universidade do Porto

[12] Sousa-Silva, R.: Computational forensic linguistics: an overview of computational applications in forensic contexts. Language and Law/Linguagem e Direito **5**(2), 118–143 (2018)

[13] Sousa-Silva, R.: Forensic linguistics: The potential of language for law enforcement in the digital age. European Law Enforcement Research Bulletin **6**, 223–232 (2022)

[14] Luhn, H.P.: Key word-in-context index for technical literature (KWIC index). American documentation **11**(4), 288–295 (1960) https://doi.org/10.1002/asi.5090110403

[15] Jeaco, S.: Helping language learners put concordance data in context. International Journal of Computer-Assisted Language Learning and Teaching **7**(2), 22–39 (2017) https://doi.org/10.4018/IJCALLT.2017040102

[16] Seretan, V., Wehrli, E.: Syntactic concordancing and multi-word expression detection. International Journal of Data Mining, Modelling and Management **5**(2), 158 (2013) https://doi.org/10.1504/IJDMMM.2013.053694

[17] Seretan, V., Wehrli, E.: Tools for syntactic concordancing. In: Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 493–500 (2010). https://doi.org/10.1109/IMCSIT.2010.5679742 . IEEE

[18] Michael, B.: Parallel concordancing and translation. In: Proceedings of Translating and the Computer 26, London, UK (2004). https://aclanthology.org/2004.tc-1.9

[19] Ohara, K.H., Fujii, S., Ohori, T., Suzuki, R., Saito, H., Ishizaki, S.: The Japanese framenet project: An introduction. In: Proceedings of LREC-04 Satellite Workshop "Building Lexical Resources from Semantically Annotated Corpora"(LREC 2004), pp. 9–11 (2004)

[20] Grady, A.: Daubert and expert testimony. AMA Journal of Ethics **8**(2), 97–100 (2006) https://doi.org/10.1001/virtualmentor.2006.8.2.hlaw1-0602

[21] Eder, M., Rybicki, J., Kestemont, M.: Stylometry with R: A package for computational text analysis (2016). https://doi.org/10.32614/RJ-2016-007

[22] Nini, A.: Idiolect: An R package for forensic authorship analysis (2024). https://cran.r-project.org/package=idiolect

[23] Millican, P.: Signature Stylometric System Version 1.0 [software]. Stylometric software for authorship attribution. http://www.philocomp.net/texts/signature.htm

[24] Juola, P.: JGAAP: A system for comparative evaluation of authorship attribution. Journal of the Chicago Colloquium on Digital Humanities and Computer Science **1**(1), 1–5 (2009)

[25] Juola, P.: Detecting stylistic deception. In: Proceedings of the Workshop on Computational Approaches to Deception Detection, pp. 91–96 (2012)

[26] Anthony, L.: Addressing the challenges of data-driven learning through corpus tool design–in conversation with laurence anthony. In: Corpora for Language Learning: Bridging the Research-practice Divide, pp. 9–18. Routledge, London (2024). https://doi.org/10.4324/9781003413301-2

[27] Scott, M.: Developing Wordsmith. International Journal of English Studies **8**(1), 95–106 (2008)

[28] Brezina, V.: Corpus linguistics and AI:#lancsbox x in the context of emerging technologies. International Journal of Language Studies **19**(2) (2025) https://doi.org/10.5281/zenodo.15250820

[29] Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlỳ, P., Suchomel, V.: The Sketch Engine. Lexicography **1**(1), 7–36 (2014) https://doi.org/10.1007/s40607-014-0009-9

[30] Gries, S.T.: Quantitative Corpus Linguistics with R: A Practical Introduction. Routledge, London (2016)

[31] Boswijk, V., Coler, M.: What is salience? Open Linguistics **6**(1), 713–722 (2020) https://doi.org/10.1515/opli-2020-0042

[32] Pojanapunya, P., Watson Todd, R.: Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. Corpus Linguistics and Linguistic Theory **14**(1), 133–167 (2018) https://doi.org/10.1515/cllt-2015-0030

[33] Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc., Sebastopol,

CA (2009)

[34] Blake, J., Tamura, K., Kredens, K.: Dual-mode n-gram similarity detection for forensic authorship analysis. In: Proceedings for 39th Pacific Asia Conference on Language, Information and Computation (PACLIC 39) (2026). https://aclanthology.org/2025.paclic-1.75.pdf

[35] Nguyen, D.T., Sat, C.G., Blake, J., Pyshkin, E.: Wikifirst: A genre-fixed, content-controlled corpus for evaluating content effects in authorship analysis. In: Proceedings of 10th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 323–327 (2026). https://aclanthology.org/anthology-files/anthology-files/pdf/latechclfl/2026.latechclfl-1.31.pdf

[36] Sat, G.C., Blake, J., Pyshkin, E.: Modelling the relative contributions of stylistic features in forensic authorship attribution. In: Proceedings for the Recent Advances in Natural Language Processing 2025 Conference, pp. 1066–1073 (2025). https://acl-bg.org/proceedings/2025/RANLP%202025/pdf/2025.ranlp-1.123.pdf

[37] Coulthard, M., Johnson, A., Wright, D.: An Introduction to Forensic Linguistics: Language in Evidence. Routledge, London (2016)