A Computational Forensic Linguistic Analysis of Narrative and Question-Answer Structures in Italian Police Interrogation Transcripts

Romane Werner,[1] **Thomas François,**[1] **Sonja Bitzer**[2]
[1]Institute for Language and Communication
[2]Institute for Interdisciplinary Research in Legal and Criminological Sciences
Université catholique de Louvain
romane.werner@uclouvain.be
thomas.francois@uclouvain.be
sonja.bitzer@uclouvain.be

Abstract

Police interrogation transcripts are key evidential documents, yet their linguistic form is rarely systematically analyzed, despite directly shaping judicial interpretation. This study presents the first computational forensic linguistic profiling of Italian police transcripts, focusing on the two transcription formats used in practice: narrative monologues and question-answer (Q-A) transcripts. Using automated extraction of 147 linguistic features, we analyze 50 authentic transcripts against a multigenre Italian reference corpus to support more transparent evaluation of police transcripts by clarifying how transcription formats systematically shape evidential interpretation in judicial contexts. Narrative monologues exhibit deeper syntactic embedding, higher past-tense usage, and more first-person singular verbs, supporting coherent and temporally ordered recounting of events. Q-A transcripts, by contrast, show longer subordinate chains, more clausal complements, and higher pronoun frequency, reflecting interactive turn-taking and procedural dynamics. Rather than aiming at predictive classification, the study reveals the linguistic mechanisms shaping transcription formats and demonstrates that structurally and legally informed features reliably distinguish them. Computational models reliably capture genre-specific cues, offering scalable, empirically grounded insights into transcription practices and evidential reliability.

1 Introduction

Police interrogation transcripts are critical evidential documents, relied upon to assess statement reliability and attribution of responsibility in criminal proceedings (Gibbons, 2003; Coulthard and Johnson, 2007). Despite their evidential centrality, transcription practices vary considerably across jurisdictions and are rarely subjected to systematic linguistic evaluation (Eades, 2010; Eerland and van Charldorp, 2022). Transcripts are often treated as neutral representations of spoken statement. However, transcription mediates between oral and written modes and between personal narrative and institutional inscription (Komter, 2006). Transformations introduced during transcription (e.g., omissions and reformulations) can directly influence how judicial actors interpret meaning, thereby affecting the suspects' credibility and intent (Shuy, 1998; Fraser, 2003). In Italy, weakly regulated transcription protocols allow police officers to reshape statements into more coherent written forms (Sinatra, 2014; Bussu, 2016), raising concerns regarding the transparency and evidential reliability of transcripts.

This study sits at the intersection of computational forensic linguistics and automatic genre analysis. The former extends computational methods to legally relevant texts, identifying systematic linguistic patterns and markers that signal interpretive significance, while the latter situates these patterns within the communicative goals and institutional conventions of the documents (Bawarshi and Reiff, 2010). Previous work has shown that features, such as POS distributions and syntactic structures reliably indicate genre-specific characteristics (Cimino et al., 2017). However, police transcripts remain largely unexplored computationally, and no prior study has systematically compared the linguistic features and the communicative functions of Italian transcripts across transcription types. To address this gap, we implement a computational pipeline encompassing automatic feature extraction and feature selection, followed by statistical testing of feature distributions. Adopting an explanatory approach, we aim to identify the linguistic variables that distinguish transcription formats and their relation to institutional and evidential functions, with classification models used solely as diagnostic tools emphasizing interpretability and forensic transparency (Branting et al., 2020).

We analyze a corpus of 50 authentic tran-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 590–603

scripts representing the two transcription types used in Italy: narrative monologues and Q-A exchanges. Our central research question is the following: Which lexical, morphosyntactic, syntactic, complexity-related, and legal features most clearly distinguish justice collaborators' interrogations by format, and how do these features compare with those typical of stereotypical genres, such as legal language, newspapers, literary texts, parliamentary discourse, and semi-structured oral interviews? We hypothesize that both monologic and Q-A transcriptions will exhibit features typical of storytelling (e.g., past-tense verbs) (Biber and Conrad, 2009; Semino and Mick, 2004). Monologic transcriptions are expected to combine these features with markers of formal legal discourse (e.g., specialized vocabulary), reflecting their role as "institutionally refracted narratives" (O'Toole, 2018). Q-A transcriptions, by contrast, are anticipated to primarily display features characteristic of spontaneous oral interaction, while also incorporating salient legal discourse markers (Drew and Heritage, 1992).

The paper is organized as follows: Section 2 reviews prior work on transcription practices and automatic genre analysis; Section 3 describes the corpus; Section 4 introduces the methods; Section 5 presents the results comparing transcription formats with reference genres; Section 6 discusses the findings; and Section 7 concludes, highlighting limitations and directions for future research.

2 Related work

The present analysis focuses on two transcription types, namely narrative monologues and Q-A exchanges, which, though coexisting within investigative discourse, are represented through distinct forms. Cross-national research in the UK, the Netherlands, and Sweden shows that transcription format can influence evidential interpretation and perceived credibility, underscoring the functional significance of structural choices (Richardson et al., 2022; Komter, 2022). In Italy, weakly regulated transcription protocols often result in the literarization of speech in narrative monologues and a preservation of dialogic elements in Q-A transcripts. Narrative monologues, typically produced in the absence of audio or video recordings, present interviewees' statements as first-person narratives (Bussu, 2016). Although ostensibly authored by the interviewee, these texts are composed by the interviewer, with question omitted, transforming the dialogic exchange into a linear written account that prioritizes readability and narrative cohesion over interactional detail (Sinatra, 2014; Bussu, 2016). Consequently, the original dialogic structure and pragmatic conditions of elicitation are obscured. By contrast, Q-A transcripts, sometimes termed *verbatim*, preserve speaker turns and interactional dynamics, providing a closer representation of the original speech (Bussu, 2016).

Computational genre analysis offers a framework for linking linguistic form to communicative purpose and institutional function (Stamatatos et al., 2000; Bawarshi and Reiff, 2010; Bhatia, 1993). Research on Italian corpora demonstrates that syntactic structures, POS distributions, lexical accessibility, and readability metrics vary systematically across literary, journalistic and legal genres (Dell'Orletta et al., 2013; Venturi, 2012; Brunato, 2014). Literary texts typically exhibit high verb and pronoun usage, reflecting narrative dynamism, whereas legal texts are characterized by nominal density, syntactic rigidity and specialized vocabulary, reflecting formal and informational goals (Biber and Conrad, 2009). A genre-based approach captures systematic linguistic patterns that encode interactional dynamics and narrative organization. Computationally, these patterns can be quantified and modeled, enabling automatic genre classification. Applied to police transcripts, this approach frames narrative monologues and Q-A exchanges as functionally motivated forms rather than mere stylistic variants, highlighting how genre-specific structures in transcription type shape both communicative and evidential purposes (Komter, 2019).

Building on this functional view of genre, automatic genre classification thus provides a computational framework for modeling how linguistic features encode the communicative functions that distinguish text types (Dömötör et al., 2022). Automatic genre classification approaches typically draw on combinations of lexical, morphosyntactic, structural, and discourse-level indicators to discriminate among genres (Dömötör et al., 2022; Santini, 2007; Albi, 2013), making it particularly valuable for computational forensic linguistics, where the goal is to identify systematic patterns arising from institutional constraints and production protocols. Police transcripts are especially amenable to such modeling: despite internal diversity, they follow routinized communicative practices that generate stable linguistic features. The combination of standardized legal characteristics with distinct struc-

tural subtypes, such as narrative monologues versus Q-A formats, produces systematic variation that automatic genre classification methods can reliably capture (Albi, 2013).

Computational genre profiling builds on automatic genre classification to analyze internal variation, highlight prototypical structures and relate linguistic patterns to communicative function. In Italy, computational genre profiling has shown that POS, syntactic complexity, and lexical accessibility provide reliable discriminators across literary and scientific texts, newspaper articles and legal genres (Dell'Orletta et al., 2013; Brunato and Dell'Orletta, 2017; Cocciu et al., 2018; Cimino et al., 2017). Specifically, Dell'Orletta et al. (2013) and Venturi (2012) showed that literary texts have more verbs and pronouns and shorter sentences, while legal texts display high nominal density, longer sentences and specialized vocabulary. Non-lexical features, particularly syntactic complexity, dependency structures, and readability measures, often outperform purely lexical cues in specialized domains where texts integrate institutional constraints with narrative elements (Dell'Orletta et al., 2014; Stamatatos et al., 2000). Together, these findings provide a robust empirical foundation for computationally modeling police transcription formats, enabling the systematic identification of format-specific linguistic patterns. While prior studies have examined Italian legal corpora or narrative texts separately, no research has yet applied these approaches to characterize Italian police transcripts, motivating the present study.

3 Data

To address our research question, we compiled a corpus integrating authentic Italian police interrogation transcripts with a multi-genre reference corpus to support systematic comparative analysis. The primary dataset comprises interrogations with justice collaborators associated with *Cosa Nostra*, obtained from the publicly accessible archives of the Italian Antimafia Commission[1]. Two distinct transcription formats are used in Italy: narrative monologues and Q-A transcripts. We collected 25 transcripts of each type, covering the period 1984-2018, resulting in 50 texts, totaling 259,067 tokens, with 46,810 tokens from monologic transcripts and 212,257 tokens from Q-A transcripts[2].

To contextualize the linguistic features of police transcripts and to support genre-based comparisons, we compiled a reference corpus representing five stereotypical genres (see Table 1): legal-lay texts, newspaper articles, semi-structured interviews, literary prose, and parliamentary discourse. Legal-lay texts were drawn from the Italian subcorpus of CorIELLS (Busso, 2021); newspaper articles were randomly sampled from *Il Fatto Quotidiano*; semi-structured interviews were sourced from the HABLA corpus (Schmidt and Wörner, 2012); literary prose contains eight contemporary novels (see Appendix C); and parliamentary discourse was included from the ParlaMint corpus, a multilingual and comparable collection of parliamentary debates from 29 European countries (Erjavec, 2024).

All corpora[3] were processed using a suite of NLP tools, which have proven effective for both general and legal Italian texts (Venturi, 2012). Morphosyntactic annotation was performed with the FDO-POS tagger (Dell'Orletta, 2009), and syntactic parsing employed DeSR (Attardi, 2006), generating CoNLL-compatible dependency formats for subsequent analysis. These tools are particularly suitable for our corpora as they have been optimized on legal Italian (TEMIS corpus) (Venturi, 2012) and validated through a manual evaluation: the POS tagger achieved precision of 0.97 on narrative texts and 0.98 on question-answer transcripts. Most remaining errors were due to dialectal or non-standard lexical items underrepresented in standard training data. This combination of domain adaptation and empirically verified accuracy ensures that the linguistic annotations are robust, reliable and representative, supporting the extraction of meaningful morphosyntactic and syntactic features for analysis.

4 Methods

4.1 Genre Feature Extraction

Genre feature extraction aimed to identify the most salient linguistic characteristics across police transcripts and the multi-genre reference corpus, supporting detailed genre analysis and providing features suitable for the automatic analysis of transcription types. To capture complementary dimensions of linguistic form, we adopted a multi-subset

[1] https://www.archivioantimafia.org/

[2] Access to Italian interrogation transcripts is legally restricted. The corpus contains the full set of publicly available material from the Antimafia Commission Archives. Despite its small size, it holds exceptional evidential value, representing rare institutional texts otherwise inaccessible for research.

[3] Currently not publicly available due to university ownership.

Text Type	Period	#Tokens	#Texts
Narrative mon.	1984-2018	46,810	25
Q-A format	2008-2009	212,257	25
Legal texts	2019	260,127	25
Newspaper articles	2015	235,987	25
Semi-structured int.	2012	149,566	25
Literary texts	1992-2009	440,158	25
Parliamentary dis.	2013	253,948	25
Total		1,598,853	175

Table 1: Overview of the corpus composition

approach, combining semi-automatic and fully automatic methods, consistent with established computational genre analysis practices. This strategy balances interpretability with analytical depth.

The genre feature set comprises 147 variables drawn from three complementary sources (see Appendix B). The first subset includes 13 semi-automatically annotated lexical and morphosyntactic features recurrent in Italian legal discourse (i.e., imperfect tense, abbreviations, participial nouns/adjectives, dialogism, anteposition patterns, enclisis, modal constructions, derivational suffixes in -ità and -(t)ivo/-(t)orio, use of present participles with verbal value, technical terms), which were informed by prior research on Italian legal discourse (Pianese, 2008; Ondelli, 2014; Visctonti, 2010; Masa and Pandimiglio, 2020). Each feature had explicit rules, for example: only participial adjectives counted for the "participial adjective" feature, and only nouns ending in -ità with legal meaning counted for the "-ità derivational nouns" feature. A Python-based semi-automatic annotation procedure was implemented: the system pre-annotated candidate instances of these phenomena, which were then verified and corrected by a trained annotator following detailed linguistic guidelines derived from prior studies. Candidate instances were first identified by a Python-based pre-annotation script and then manually verified by a single trained annotator, who is the only person responsible for the annotation. To ensure reliability despite the single annotator, the semi-automatic procedure was quantitatively evaluated over all annotated instances, achieving Precision = 0.86, Recall = 0.90, and F1 = 0.88 (micro-averaged across features), providing an objective measure of annotation quality.

The second subset comprises nine readability and lexical accessibility metrics from the READ-IT framework (Brunato et al., 2020), including the global readability index and its subcomponents, the Gulpease index (Lucisano and Piemontese, 1988), and VDB[4]-based lexical sophistication measures. These indicators capture text complexity dimensions known to differentiate legal, journalistic, and literary genres (Dell'Orletta et al., 2013; Brunato, 2014; Cocciu et al., 2018). The third subset consists of 125 morphosyntactic and syntactic features automatically extracted with Profiling-UD (Dell'Orletta et al., 2013), covering dependency relations, clause embedding, structural complexity, and syntactic variability. Integrating these three subsets ensures comprehensive coverage of both legally distinctive markers and genre-relevant structural patterns, yielding a robust feature space for modeling transcription-specific linguistic profiles. Overall, this multi-level extraction yielded 147 features, combining domain-specific, readability and syntactic dimensions, providing a comprehensive linguistic representation.

4.2 Feature selection

We implemented a feature selection strategy to address common challenges in high-dimensional linguistic datasets (e.g., overfitting, limited generalizability, and reduced interpretability) (Bouchlaghem et al., 2022; Tang et al., 2014). Feature selection identifies the most informative variables, removing noise and redundancy while enhancing interpretability, a key requirement in computational analyses of legal texts (Bouchlaghem et al., 2022).

Feature selection methods generally fall into three broad categories. Filter methods evaluate features independently of predictive model, using metrics such as information gain (Tang et al., 2014). Wrapper methods assess feature subsets by iteratively training and evaluating models to measure the predictive contribution of candidate feature sets (e.g., Recursive Feature Elimination) (Tang et al., 2014). Embedded methods integrate feature selection within model training, as in L1 or L2 regularization (Tang et al., 2014).

Following best practices in computational linguistics and automatic genre classification (Cai et al., 2018), we implemented a two-stage hybrid pipeline combining filter and wrapper strategies. First, SelectKBest (filter-based) with the ANOVA F-test (f-classif) ranked features by individual discriminative power. The top k features were selected based on cross-validated discriminative per-

[4]*Vocabolario di Base*, i.e., basic vocabulary.

formance, retaining variables that maximized predictive utility while minimizing redundancy. Second, the reduced set was refined using RFECV with 10-fold cross-validation and an ExtraTreesClassifier estimator, enabling detection of feature interactions. This sequential design balances computational efficiency with the identification of both individually salient and combinatorial feature effects (Cimino et al., 2017).

Feature selection was applied jointly over the two transcription types and the multi-genre reference corpus, yielding variables that discriminate between transcription formats and against other genres. From the original 147 features, 52 were retained (63.8% reduction), dominated by syntactic indicators (23/72), readability metrics (5/9), and legal-lexical markers (8/13). General lexical variety contributed minimally, indicating that structural and domain-specific features provide the strongest basis for forensic genre analysis.

Classifiers trained on the selected features with an 80/20 train-test split achieved stable performance. Models using only SelectKBest features performed comparably to RFECV-refined models (accuracy = 0.63; macro F1 = 0.75)[5]. SelectKBest captured strong individual effects, while RFECV highlighted interaction-sensitive patterns, demonstrating the complementary value of the two approaches.

We use feature-based models because they provide a direct and transparent mapping from linguistic features to outcomes, allow explicit quantification of feature effects and are more stable for the moderate dataset size used here. While transformer-based models could be analyzed with post-hoc interpretability tools such as SHAP, feature-based models offer a more straightforward and interpretable approach for revealing systematic patterns in transcription formats.

4.3 Statistical Testing

This study employs a three-stage analytical pipeline to identify the most discriminative linguistic features across text types, integrating (i) discrimination scores, (ii) statistical hypothesis testing, and (iii) effect-size estimation. Together, these steps ensure that selected features are statistically reliable and linguistically interpretable.

First, discrimination scores quantify univariate deviations across groups following Guyon and Elisseeff (2003). For each feature and target group, the score measures the difference between the feature's mean in the target group and the mean in all other groups, divided by the pooled standard deviation. Larger absolute scores indicate greater divergence of the target group from the others. Positive values indicate over-representation in the target group, while negative values indicate under-representation. Because scores are standardized, they allow direct comparison across heterogeneous features. All values were cross-validated against descriptive statistics to ensure consistency with the underlying distributions.

Second, we assessed feature reliability through hypothesis testing. Normality and homogeneity of variance were evaluated using Shapiro-Wilk and Levene tests. Parametric tests (independent-samples t-test for pairwise contrasts; ANOVA for multi-group comparisons) were applied when assumptions were met, otherwise non-parametric alternatives (Mann-Whitney U; Kruskal-Wallis) were used. This ensures robust inference across features with differing distributional properties.

Finally, effect sizes were computed using Pearson's r, a standard measure in computational linguistics and genre-classification research. This metric quantifies the strength of association between each feature and group membership, providing a standardized estimate of discriminability and facilitating systematic interpretation of linguistic variation.

4.4 Ensemble Analyses

This study examines transcription types against multiple genres to identify both the general and distinctive linguistic properties of police interrogation transcripts. This ensemble design is analytical rather than predictive, moving from broad transcriptional tendencies toward fine-grained and format-specific distinctions. To capture variation across different levels of granularity, we developed a four-part ensemble framework, each configuration addressing a complementary analytical perspective (Chen and Kubát, 2025):

1. Ensemble 1: Both transcription types are compared collectively against the whole reference corpus, isolating features that characterize

[5]Although certain linguistic features occur more frequently in some transcription types, this distributional variation reflects meaningful differences across types rather than a dataset flaw. The analysis is designed to capture these systematic patterns, and macro-F1 is reported to provide a fair evaluation across all classes, ensuring that features characteristic of less frequent patterns are properly accounted for.

transcriptional language as a whole.

2. Ensemble 2: Each transcription type (i.e., monologic narratives and Q-A format) is compared individually against all other genres, including the other transcription type, highlighting linguistic features distinctive of each transcriptional format.

3. Ensemble 3: The two transcription types are evaluated jointly with each genre individually, identifying shared transcriptional tendencies and divergences across communicative domains.

4. Ensemble 4: Each transcription type is compared separately with each reference genre and with the other transcription type, offering maximal granularity for detecting genre-specific linguistic patterns.

These four ensembles form a hierarchical design that moves from broad transcriptional generalizations (Ensemble 1) to detailed cross-genre contrasts (Ensemble 4), supporting both macro-level comparability and micro-level interpretability.

This study applies established computational methods to the specialized domain of police interrogation transcriptions, integrating three subsets of linguistic representation, namely legal-specific features, readability metrics and detailed syntactic measures, within a unified analytical framework. By combining filter- and wrapper-based feature selection with hierarchical ensemble comparison, the approach moves beyond lexical statistics to capture structural and institutional cues relevant to the transcription context. Structurally and legally informed features are expected to provide greater discriminative power than purely lexical features, illustrating the value of domain-aware modeling in legal NLP.

5 Results

5.1 Syntactic Complexity and Structural Embedding

Both transcription types exhibit notably higher syntactic complexity than most written reference corpora. Q-A transcripts average 4.30 verbal heads per sentence[6], substantially exceeding literary texts (1.44, $p < 0.001$, r = -0.82) and legal texts (2.46, $p < 0.001$, r = -0.67). This reflects the interactional demands of the Q-A format, where multiple predicates are often embedded within a single turn (see Appendix 6). Narrative monologues show slightly lower values (3.60) but still surpass most reference genres ($p < 0.001$), consistent with extended narrative sequences that support sequential event presentation and temporal coherence. Together, these patterns indicate that both transcription types favor multi-predicate sentence structures (see Table 2; Appendix A for an explanation of all the variables).

Sentence length mirrors this syntactic density. Q-A transcripts average 32.81 tokens per sentence, exceeding parliamentary discourse (27.12, $p = 0.01$, r = -0.33) and newspaper articles (10.58, $p < 0.01$, r = -0.84), reflecting dense information packaging within interactive turns. Narrative monologues show comparable values (31.54 tokens), differing minimally from Q-A transcripts ($p = 0.60$, r = -0.07) but remaining significantly longer than parliamentary discourse ($p < 0.05$, r = -0.36) and newspapers (24.08, $p < 0.001$, r = -0.61). Their similarity to legal texts (30.68, $p = 0.56$, r = -0.08) and divergence from semi-structured interviews (11.59, $p < 0.01$, r = -0.89) suggests that monologues preserve extended syntactic units to support detailed event narration.

Measures of hierarchical structure reveal further contrasts. In Q-A transcripts, clausal complement distance[7] (1.46) and subordinate chain length (1.52) are substantially higher than in parliamentary discourse (0.77 and 1.29, respectively; both $p < 0.001$, $r < -0.81$) and newspaper writing (0.74 and 1.19, respectively; both $p < 0.001$, $r < -0.79$) (see Appendix 3 and Figure 1), indicating a tendency toward nested, multi-layered syntax in interactive exchanges. Narrative monologues show slightly lower values (1.36 for clausal complements) but still exceed literary texts in subordinate chain length (1.19; $p < 0.001$, r = -0.65), reflecting the structural organization required for temporally sequenced storytelling.

5.2 Verbal Activity and Narrative Orientation

Temporal and verbal patterns clearly distinguish narrative monologues from both Q-A transcripts and conventional written genres. Monologues exhibit a notably high proportion of verbal roots (89.10%) and past-tense verbs (57.96%), reflecting

[6] e.g., *lui diceva che voleva andare a vedere se riusciva a trovare qualcosa.*; EN trad. *He said he wanted to go and see if he could find something.*

[7] e.g., *"...gli diceva il dottor La Barbera: a Luigi, tuo fratello, lo hanno ammazzato gli Scarantino..."*.; EN trad. *Dr "... La Barbera told him: Luigi, your brother, was killed by the Scarantino family..."*.

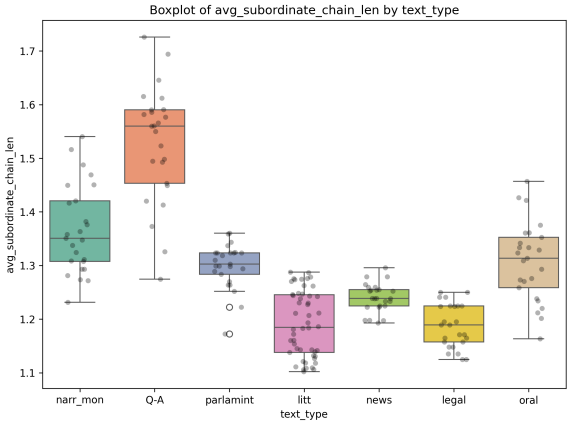

Figure 1: Distribution of avg_subordinate_chain_len across transcription types.

their event-oriented narrative structure. These values exceed those observed in literary prose (65.55% verbal roots, $p < 0.001$, r = -0.71; 40.67% past tense, $p < 0.001$, r = -0.59) and semi-structured interviews (67.44% verbal roots, $p < 0.001$, r = -0.89; 29.90% past tense, $p < 0.001$, r = -0.90). Legal texts also show elevated past-tense usage (62.90%, $p < 0.01$, $r = 0.90$), likely reflecting codified reporting conventions rather than narrative recounting. Overall, the tense profile of monologues highlights their reliance on sequential past-event narration and dense verbal predication to maintain temporal coherence.

First-person singular verbs further mark the narrative perspective of monologues[8]. Monologues average 26.30 occurrences, exceeding Q-A transcripts (21.91) and literary texts (14.12, $p < 0.05$, r = -0.49), underscoring strong self-anchoring typical of testimonial narration. Q-A transcripts also favor first-person singular forms relative to literary texts ($p < 0.05$, r = -0.38), reflecting the inherently subjective stance of interviewer-interviewee exchanges. Semi-structured interviews (22.39) align closely with Q-A transcripts (p = -0.68, $r = 0.05$), consistent with their more collaborative and dialogic orientation.

Imperfect-tense usage further characterizes narrative structuring. Monologues include 18.39% imperfect forms, supporting their function in expressing habitual past actions. While not significantly different from Q-A transcripts (15.24, p = -0.16, $r = 0.20$), monologues diverge from literary texts (19.78, $p = 0.03$, $r = 0.24$) and sharply from semi-structured interviews (5.92, $p < 0.05$, r = -0.59). This positions monologues as an intermediate narrative format: higher than in interactive interviews but slightly lower than in literary exposition. Q-A transcripts exhibit significantly higher imperfect usage than interviews ($p < 0.05$, r = -0.76), reflecting real-time recounting of ongoing actions within interaction.

Finally, first-person plural usage highlights the interactional orientation of Q-A transcripts (see Appendix 5). With an average of 9.74% Q-A exchanges employ plural forms significantly more frequently than narrative monologues (6.19, p = -0.02, $r = 0.30$), indicating a stronger reliance on jointly framed perspectives and shared participation structures. Although parliamentary discourse displays even higher rates (14.90, $p < 0.05$, $r = 0.60$), Q-A transcripts nonetheless align more closely with collaborative, multi-party discourse practices than with the predominantly individual viewpoint of monologic narration.

5.3 Interactional Features and information-structuring strategies

Pronouns and object preposing reveal distinct interactional and pronominal strategies across transcription types. Q-A transcripts exhibit a high pronoun frequency (10.20), substantially exceeding parliamentary discourse (4.77, $p < 0.001$, r = -0.93) and legal texts (2.57, $p < 0.001$, r = -0.97), reflecting the centrality of interlocutor reference and turn-taking (see Figure 2). Narrative monologues show moderate pronoun use (6.73), consistent with self-focused storytelling, and differ significantly from both Q-A transcripts ($p < 0.001$, r = -0.83) and most written genres.

Object preposing is particularly prominent in Q-A transcripts (50.14), contrasting sharply with newspaper articles (35.66, $p < 0.001$, r = -0.97) and legal texts (5.55, $p < 0.001$, $r = 0.99$), supporting the foregrounding of key entities in interactive Q-A sequences. Monologues occupy an intermediate position (29.81), significantly lower than Q-A transcripts ($p < 0.001$, r = -0.70) but higher than most written reference texts, linking structural emphasis with narrative function.

6 Discussion

Prior studies in automatic genre analysis provide a valuable benchmark for interpreting our results. Literary texts typically exhibit high proportions of

[8] e.g., *Ho conosciuto Spatuzza per il tramite di Giuseppe Graviano...* EN trad. *I met Spatuzza through Giuseppe Graviano...*

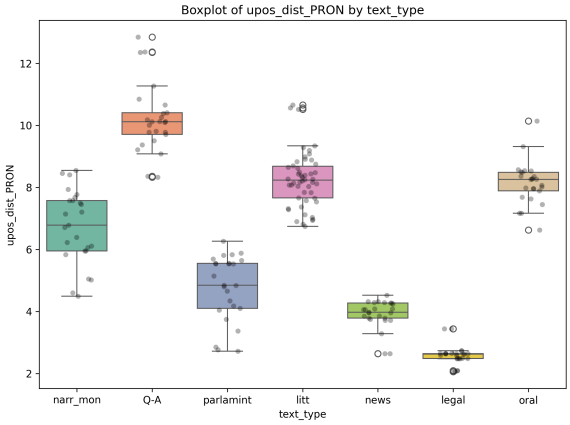

Figure 2: Pronoun frequency (`upos_dist_PRON`) across transcription types.

Feature	Transc.	Ref.	p-value	ES
verbal-head-per-sent	3.60/4.36	1.44–2.67	***	.62–.90
tokens-per-sent	31.54/32.81	10.58–30.58	**	.08–.89
dep-dist-ccomp	1.37/1.46	0.23–1.29	**	.11–.94
avg-sub-chain-len	1.36/1.52	1.18–1.30	**	.36–.90
verbal-root-perc	89.12/76.21	65.55–81.09	***	.12–.89
verbs-tense-dist-Past	57.96/43.62	29.90–62.50	***	.15–.93
verbs-num-pers-Sing1	26.30/21.91	2.09–22.39	***	.23–.94
verbs-tense-dist-Imp	18.39/15.24	0.62–19.78	***	.06–.93
verbs-num-pers-Plur1	6.19/9.74	0.27–14.90	***)	.10–.82
upos-dist-PRON	6.73/10.20	2.57–8.24	***	0.58–0.97
obj-pre	29.81/50.14	5.55–44.73	*	0.28–0.99

Table 2: Descriptive statistics and test results for each feature. *Transc.: narrative/Q–A means. Ref.: min–max across reference genres. p: significance (* $p<0.05$, ** $p<0.01$, *** $p<0.001$). ES = effect size.*

verbs and pronouns, reflecting interactional and narrative discourse (Dell'Orletta et al., 2013; Biber and Conrad, 2009). Our findings confirm this pattern: both transcription types exhibit elevated verbal and pronominal rates, aligning them with literary narratives and semi-structured interviews, while nouns are comparatively less frequent. These distributions reveal the fundamentally dialogic nature of police interrogations, in which utterances are co-constructed and contextually responsive. By contrast, legal corpora prioritize nouns, adjectives, and numerals (Venturi, 2012; Brunato, 2014), a pattern replicated in our data (nouns = 29.95%).

Sentence-level measures further differentiate transcripts from other genres. While literary texts exhibit relatively short sentences (Dell'Orletta et al., 2013), police transcripts, considered across both transcription formats, show long average sentence lengths (32.17 words), comparable to legal texts (Venturi, 2012: 30.13). This density likely reflects spontaneous elaboration, clause embedding, and the cognitive effort of reconstructing events under questioning. Syntactic flexibility also contrasts with legal texts, which maintain canonical object positioning in over 94% of cases; in police transcripts, only 55-58% of sentences follow this norm, with monologues slightly closer to the legal standard. This variation reflects the dialogic negotiation of information in Q-A exchanges, where speakers foreground key referents in response to prompts.

Narrative monologues are characterized by high verbal activity, first-person perspective, and past-tense narration, supporting coherent temporal and causal accounts of events. These features serve testimonial purposes, enabling timeline reconstruction, plausibility assessments, and narrative-coherence evaluation (Triki, 2023), though their complexity may obscure details. By contrast, Q-A transcripts display interactional linguistic patterns, including elevated pronoun density, complex syntax, and object preposing, reflecting the responsive nature of interrogation (Haworth, 2010). They capture not only the interviewee's answers but also the questioning practices that shape them, including reformulation and leading questions (Shuy, 1998). Their communicative purpose is therefore investigative and dialogic: to elicit information, clarify referents, and shape the trajectory of statement. From an evidential standpoint, the Q-A structure enhances procedural transparency which allows to assess how particular answers emerged or whether they were influenced by questioning style.

The coexistence of Q-A transcripts and narrative monologues demonstrates that the notion of a "neutral" transcript is a legal fiction: each format carries distinct communicative and evidential functions that shape how meaning, responsibility, and credibility are represented in transcripts. Misinterpreting Q-A exchanges as equivalent to narrative accounts risks biased credibility assessments, while treating monologues as unmediated truth ignores their co-construction within institutional constraints. Both formats instantiate the dialogic tension between personal narration and institutional inscription, in which linguistic evidence is constructed rather than merely recorded.

Recognizing the distinct affordances of these transcription types has direct implications for investigative and judicial practice (Haworth, 2010; Richardson et al., 2022). Together, these formats function as complementary forensic tools, illustrating that linguistic form, communicative purpose, and evidential function are inseparable. Far from being mere written records of speech, police inter-

rogation transcripts are hybrid institutional genres that integrate narrative reconstruction, spontaneous interaction, and legal codification, shaping how evidence is framed and how accountability, agency, and credibility are represented from the interview room to the courtroom.

These differences are not merely linguistic but carry direct evidential implications. Differences in syntactic embedding, pronominal structure, and clause chaining influence how statements are interpreted in judicial settings, affecting assessments of credibility, voluntariness, and the extent to which answers may have been shaped by questioning practices. Computational methods that isolate these patterns provide a principled basis for evaluating how transcription formats mediate representation, supporting more transparent and informed forensic interpretation.

For legal NLP, our findings underscore the importance of genre-sensitive modeling. Treating transcripts as undifferentiated legal text risks not only misclassification but also misinterpretation of how statements are produced and how evidential meaning is constructed. Incorporating genre-specific linguistic cues, such as pronoun density, clause-chaining patterns, enables models to capture the procedural and dialogic conditions under which statements emerge. Computational analysis thus goes beyond descriptive classification, providing an empirically grounded framework for tracing how discourse practices shape evidential content and supporting more transparent, consistent, and accurate forensic evaluation. By modeling these linguistic patterns, legal NLP can help analysts, lawyers, and judges assess statements more reliably and fairly.

7 Conclusion

This study examined two transcription formats used in Italian police interrogations, namely narrative monologues and Q-A transcripts, through a computational forensic linguistics and genre analysis lens. The results demonstrate that these formats constitute distinct yet complementary hybrid genres, each encoding different communicative and evidential functions.

Using a multi-layered computational framework combining syntactic profiling, readability metrics, and legal-lexical features, we identified the linguistic mechanisms differentiating the two formats. Narrative monologues show greater syntactic embedding, dense verbal predication, and strong past-tense orientation, supporting temporally structured, first-person accounts of events. Q-A transcripts, in contrast, exhibit elevated object preposing, pronominal reference, and multi-layered clause chains, reflecting interactional responsiveness and the procedural logic of elicitation.

Both formats share high verb and pronoun density, aligning with oral and narrative discourse while diverging from the nominal and syntactically rigid profile of legal texts. These patterns reliably mark transcription format, yet treating them as interchangeable risks misinterpreting a suspect's credibility, intention, or the procedural context. Narrative monologues facilitate coherent event reconstruction but may obscure interviewer influence, whereas Q-A transcripts enhance procedural transparency at the expense of narrative cohesion.

Beyond theoretical contributions, this study highlights the practical value of computational genre analysis in legal NLP. Structurally and legally informed features provide a transparent and forensically meaningful basis for distinguishing transcription formats, and computational models can reliably capture genre-specific linguistic cues. The framework is reproducible, scalable, and empirically grounded, enabling systematic detection of transcriptional distortions that may affect evidential interpretation and judicial decision-making.

Overall, this study demonstrates that linguistic form, communicative purpose, and evidential function are inseparable in police transcripts. Recognizing their hybrid nature allows computational models to capture how procedural and dialogic conditions shape testimony, providing insights that enhance transparency, reliability, and fairness in forensic and judicial practice.

Limitations

This study has several limitations: the corpus size is restricted, multimodal aspects of interaction (e.g., prosody, pausing, timing) are not captured, and the analysis is limited to Italian and two transcription formats. Future research should expand the corpus, integrate multimodal signals, and explore applications for automatic detection of transcriptional mediation in institutional records, while modeling variation across interviewers, jurisdictions, and transcription protocols.

References

Anabel Borja Albi. 2013. A genre analysis approach to the study of the translation of court documents. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 12.

Giuseppe Attardi. 2006. Experiments with a multilingual non-projective dependency parser. *Proceedings of the Tenth Conference on Computational Natural Language Learning*, pages 166–170.

Anis Bawarshi and Mary Reiff. 2010. *Genre: An Introduction to History, Theory, Research and Pedagogy*. Parlor Press, Anderson.

Vijay Bhatia. 1993. *Analysing Genre: Language Use in Professional Settings*. Longman, London.

Douglas Biber and Susan Conrad. 2009. *Register, Genre and Style*. Cambridge University Press, Cambridge.

Younes Bouchlaghem, Yassine Akhiat, and Souad Amjad. 2022. Feature selection: A review and comparative study. *10th International Conference on Innovation, Modern Applied Sciences and Environmental Studies*, 351.

Karl Branting, Craig Pfeifer, Bradford Brown, Lisa Ferro, John Aberdeen, Brandy Weiss, Mark Pfaff, and Bill Liao. 2020. Scalable and explainable legal prediction. *Artificial Intelligence and Law*, 22(2):213–238.

Dominique Brunato. 2014. Complessità necessaria o stereotipi del 'burocratese'? un'indagine sulla leggibilità del linguaggio amministrativo da una prospettiva linguistico-computazionale. *XIII Congresso della SILFI*.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2020. Profiling-UD: A tool for linguistic profiling of texts. *Proceedings of the 12th Conference on Language Resources and Evaluation*, pages 7145–7151.

Dominique Brunato and Felice Dell'Orletta. 2017. On the order of words in italian: A survey on genre vs complexity. *Proceedings of the Fourt International Conference on Dependency Linguistics*.

Lucia Busso. 2021. Lexicon and grammar in legallay language: A quantitative corpus study on italian. *Studi Italiani di Linguistica Teorica e Applicata*, 51:5–32.

Anna Bussu. 2016. Gathering evidence: Problems, training requirements, and good practices in the italian judicial police force. *Police Practice and Research*, 17(5):394–407.

Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang. 2018. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79.

Xinying Chen and Miroslav Kubát. 2025. Genre variation in dependency types: A two-level genre analysis using the czech national corpus. *Proceedings of the Eighth International Conference on Dependency Linguistics*, pages 84–92.

Andrea Cimino, Martijn Wieling, Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2017. Identifying predictive features for textual genre classification: the key role of syntax. *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017*.

Eleonora Cocciu, Dominique Brunato, Giulia Venturi, and Felice Dell'Orletta. 2018. Gender and genre linguistic profiling: A case study on female and male journalistic and diary prose. *Proceedings of the Fifth Italian Conference on Computational Linguistics*.

Malcolm Coulthard and Alison Johnson. 2007. *An Introduction to Forensic Linguistics: Language in Evidence*. Routledge, London.

Felice Dell'Orletta. 2009. Ensemble system for part-of-speech tagging. *Proceedings of EVALITA*, 9:1–8.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2013. Linguistic profiling of texts across textual gernes and readability levels: An exploratory study on italian fictional prose. *Proceedings of Recent Advances in Natural Language Processing*, 26(4):471–495.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2014. Assessing document and sentence readability in less resourced languages and across textual genres. *ITL - International Journal of Applied Linguistics*, 165(2):163–193.

Paul Drew and John Heritage. 1992. Analysing talk at work: An introduction. In Drew Paul and Heritage John, editors, *Talk at Work: Interaction in Institutional Settings*, pages 3–65. Cambridge University Press, Cambridge.

Andrea Dömötör, Tibor Kákonyi1, and Zijian Győzó Yang. 2022. What's your style? automatic genre identification with neural network. *Computación y Sistemas*, 26(3):1293–1299.

Diana Eades. 2010. Verbatim courtroom transcripts and discourse analysis. In Hannes Kniffka, editor, *Recent Developments in Forensic Linguistics*, pages 241–254. Peter Lang, Frankfurt am Main.

Anita Eerland and Tessa van Charldorp. 2022. The influence of police reporting styles on the processing of crime related information. *Frontiers in Communication*, 7.

Tomaz Erjavec. 2024. Multilingual comparable corpora of parliamentary debates in Parlamint 4.1.

Helen Fraser. 2003. Issues in transcription: Factors affecting the reliability of transcripts as evidence in legal cases. *Forensic Linguistics*, 10(2):203–226.

John Gibbons. 2003. *Forensic Linguistics: An Introduction to Language in the Justice System*. Wiley, Hoboken.

Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Kate Haworth. 2010. Police interviews in the judicial process: Police interviews as evidence. In Coulthard Malcolm and Johnson Alison, editors, *The Routledge Handbook of Linguistics*, pages 241–254. Routledge, Abingdon.

Martha Komter. 2006. From talk to text: The interactional construction of a police record. *Research on Language and Social Interaction*, 39(3):201–228.

Martha Komter. 2019. *The Suspect's Statement: Talk and Text in the Criminal Process*. Cambridge University Press, Cambridge.

Martha Komter. 2022. Institutional and academic transcripts of police interrogations. *Frontiers in Communication*, 7.

Pietro Lucisano and Maria Emanuela Piemontese. 1988. Gulpease: una formula per la predizione della leggibilità di testi in lingua italiana. *Scuola e Città*, pages 110–124.

Viviana Masa and Matteo Pandimiglio. 2020. Linguistica, diritto e variazione: uno sguardo al linguaggio delle sentenze in italia. *Proceedings of Linguaggi settoriali e specialistici: sincronia, diacronia, traduzione, variazione*.

Stefano Ondelli. 2014. Ordine delle parole nell'italiano delle sentenze: alcune misurazioni su corpora elettronici. *Informatica e diritto*, 23(1):13–39.

Jacqueline O'Toole. 2018. Institutional storytelling and personal narratives: Reflecting on the "value" of narrative inquiry. *Irish Educational Studies*, 37(2):175–189.

Giovanna Pianese. 2008. *Analisi linguistica comparativa di un corpus di testi del dominio giuridico: Sentenze penali italiane e francesi a confronto*. Università degli Studi di Napoli Federico II, Napoli.

Emma Richardson, Kate Haworth, and Felicity Deamer. 2022. For the record: Questioning transcription processes in legal contexts. *Applied Linguistics*, 43(4):677–697.

Marina Santini. 2007. Automatic genre identification: Towards a flexible classification scheme. *BCS IRSG Symposium: Future Directions in Information Access*.

Thomas Schmidt and Kai Wörner. 2012. *Multilingual Corpora and Multilingual Corpus Analysis*. John Benjamins, Amsterdam.

Elena Semino and Short Mick. 2004. *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. Routledge, London.

Roger Shuy. 1998. *The Language of Confession, Interrogation and Deception*. Sage, New York.

Chiara Sinatra. 2014. Il passaggio dall'oralità alla scrittura in ambito forense e giudiziario. *Quadernos AISPI*, 4:197–212.

Efstathios Stamatatos, George Kokkinakis, and Nikos Fakotakis. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495.

Jiliang Tang, Salem Alelyani, and Huan Liu. 2014. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, pages 37–67.

Nesrine Triki. 2023. Narratorial techniques in Tunisian police and court transcripts: A forensic linguistic approach. *Text and Context*.

Giulia Venturi. 2012. Investigating legal language peculiarities across different types of italian legal texts: An nlp-based approach. *IAFL Porto 2012 Proceedings*.

Jacqueline Visctonti. 2010. *Lingua e diritto: Livelli di analisi*. LED Edizioni Universitarie, Milano.

A Features

Feature Type	Description	#
Semi-automatic legal annotations	Legal-specific morphosyntactic phenomena in Italian: imperfect tense, abbreviations, participial nouns and adjectives, dialogic markers, anteposition of adverbs/adjectives/participles, derivational suffixes (-ità, -(t)ivo/-(t)orio), enclisis with infinitive modal verbs, modal verb constructions, present participle usage, and domain-specific/technical terms.	13
READ-IT features	Readability indices and lexical coverage measures, including Global READ-IT score, READ-IT Base, Lexical, and Syntactic subcomponents, Gulpease index, and the proportion of lemmas included in the Vocabolario di Base (VDB), fundamental, high-usage, and high-availability vocabulary.	9
Profiling-UD features	Automatically extracted morphosyntactic and syntactic features, including raw text properties, lexical variety, morphosyntactic information, verbal predicate structure, parse tree metrics, constituent order, syntactic relations, and measures of subordination.	125
Total		147

Table 3: Summary of the features extracted across corpora.

B Feature Explanations

feature	Feature category	Feature explanation
tokens-per-sent	Raw text property	Average length of sentences
verbs-num-pers-Sing1	Morphosynt. info	Distribution of verbs in the 1st pers. sing.
verbs-num-pers-Plur1	Morphosynt. info	Distribution of verbs in the 1st pers. plur.
verbs-tense-Imp	Morphosynt. info	Distribution of verbs in the imperfect
verbs-tense-Past	Morphosynt. info	Distribution of verbs in the past tense
verbal-head-per-sent	Syntactic info	Average distribution of verbal heads
verbal-root-perc	Syntactic info	Average distribution of roots headed by a lemma tagged as a verb
obj-pre	Syntactic info	Distribution of objects preceding the verb
dep-dist-ccomp	Syntactic info	Average distribution of clausal complements
avg-sub-chain-len	Syntactic info	Average length of subordinate chains

Table 4: Explanation of selected features.

C Novels Included in the Reference Corpus

- *L'amore molesto* from Elena Ferrante (1992).
- *Va dove ti porta il cuore* from Susanna Tamaro (1994).
- *Mentre la mia bella dorme* from Rossana Campo (1999).
- *Accabadora* from Michela Murgia (2009).
- *L'acustica perfetta* from Daria Bignardi (2012).
- *Tre metri sopra il cielo* from Federico Moccia (1992).
- *Ti prendo e ti porto via* from Niccolò Ammaniti (1999).
- *Io uccido* from Giorgio Faletti (2002).

Table 5: List of novels included in the reference corpus.

D Boxplots

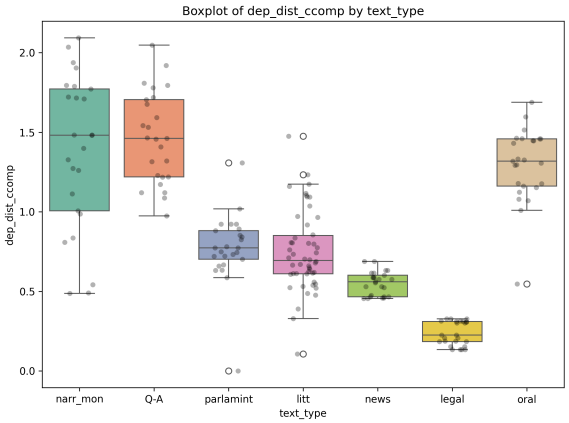

Figure 3: dep_dist_ccomp

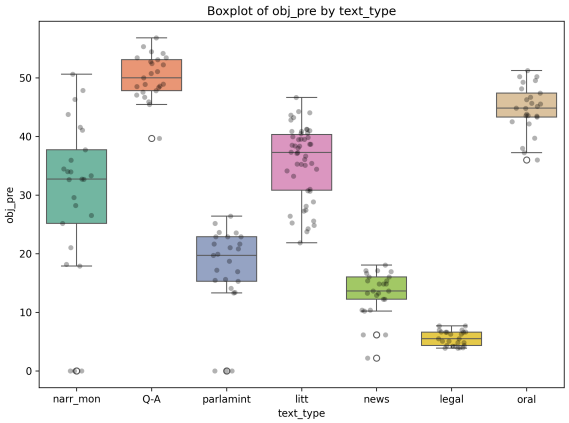

Figure 4: obj_pre

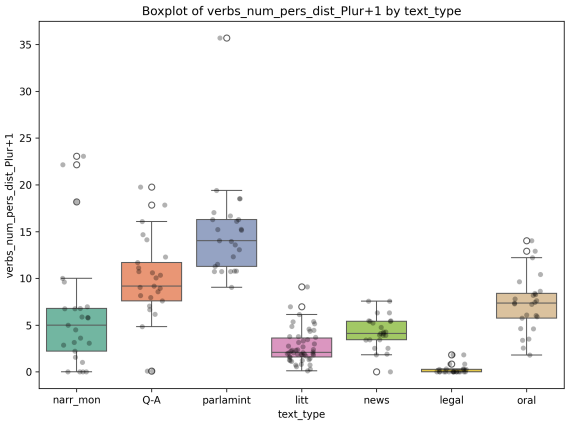

Figure 5: verbs_num_pers_dist_Plur1

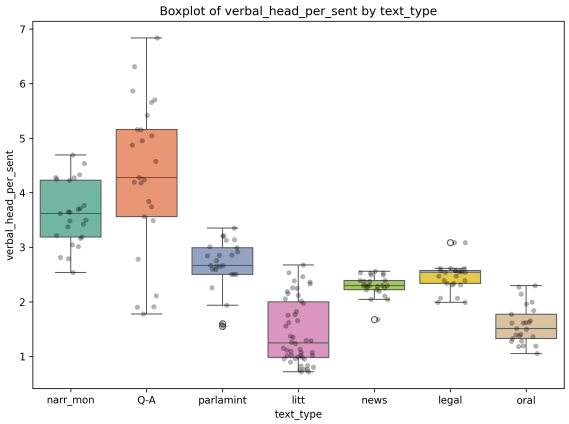

Figure 6: verbal_head_per_sent