From Decision Tree to Detection Pipeline: Formalizing van Dijk's Socio-Cognitive Framework for Automated Anti-Language Identification in RICO Transcripts

Elena Morandini
Universidad de Alicante
em126@alu.ua.es

Abstract

Automated detection of criminal language in intercepted communications remains an unsolved problem for law enforcement agencies (LEAs). Current NLP classifiers perform poorly since they assume that criminal intent is lexically explicit. Criminal organizations evade law enforcement through the systemic use of anti-language: a highly coded, context-dependent language engineered to mask conspiratorial intent beneath mundane vocabulary. This paper presents a context-first detection pipeline that reverses conventional computational architectures by establishing macro-organizational context before attempting utterance-level classification. Grounded in the socio-cognitive framework of Critical Discourse Analysis (CDA), the pipeline formalizes a seven-step analytical procedure into sequential NLP subtasks: speaker-role classification, ideological polarization detection, genre identification, move labeling, and micro-level feature extraction. The framework is evaluated on a specialized corpus of authenticated American Mafia wiretap transcripts. Feature validation reveals distinct anti-language patterns, including the absence of directive mitigation and elevated deontic saturation. By making implicit meanings explicit through a multi-layered feature vector, this pipeline offers algorithmic interpretability that meets evidentiary standards for legal expert testimony and provides a foundation for automated intelligence gathering, forensic linguistics, and cybersecurity analytics.

1 Introduction

U.S. LEAs operating under the Racketeer Influenced and Corrupt Organizations (RICO) Act, a federal law targeting organized crime through pattern-of-racketeering prosecutions, generate large volumes of wiretap transcripts that are systematically underused as evidence. The reason is computational as much as legal: NLP threat-detection systems trained on explicit, bald-on-record commissive speech acts produce high false-negative rates on the coded discourse of organized crime, where criminal intent is encoded through implicature (implied rather than explicit meaning), relexicalization (domain-specific semantic substitution, e.g., mapping "work" to "homicide"), and genre-conditioned inference rather than surface form (MacAvaney et al., 2019; Morandini, 2026). As Mafia communication moves to encrypted digital platforms, the same coded language used in historic wiretaps is now found in apps like Signal, WhatsApp, and dark-web messaging. This shift creates a significant challenge: the increasing amount of intercepted communications is outpacing the ability to analyze them effectively.

The gap is based on architectural principles. Context-free classifiers struggle to identify statements where the intent to commit a crime comes from the shared organizational context between speakers, rather than just the words they use (van Dijk, 2008, pp. 73–76; Gambetta, 2009, pp. 3–29). A phrase like "a friend of ours" serves as an organizational identifier only when evaluated within the sequential move structure of a Coded Introduction, a communicative genre that, following Giltrow and Stein (2009), functions as a pre-packaged cognitive schema shared exclusively by ingroup members. Likewise, the utterance "I'd hate to see anything happen" constitutes an implicit commissive speech act, specifically a Veiled Threat, exclusively when mapped to the Sanction move in a Sit-Down meeting. Without establishing this macro-organizational context model in advance, the pipeline prevents micro-level features from firing.

Proceedings of the Second International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security (NLPAICS'2026)
11–12 June, 2026.

This paper addresses the gap by proposing a context-first detection pipeline that employs a seven-step decision tree, applied to van Dijk's (2011) socio-cognitive CDA framework, on 18 RICO transcripts . The pipeline's core innovation is sequential, top-down architecture: Organizational context is established before individual utterances are classified, replicating the inferential logic that human analysts and criminal insiders apply naturally. The paper makes three contributions: (a) a step-by-step mapping of the decision tree to concrete NLP subtasks (see Appendix A); (b) a six-feature micro-level vector empirically validated on the RICO corpus; and (c) a specification of three operational LEA applications.

2 Related Work

The pipeline draws on three distinct research traditions that have developed largely in parallel: NLP threat detection, Forensic Linguistics, and linguistic studies of organized crime.

Current NLP systems for threat detection predominantly rely on surface-level lexical features trained on explicit hate-speech corpora (MacAvaney et al., 2019). Miró-Llinares et al. (2018) developed algorithms to detect hate speech in digital microenvironments, while Ramírez Sánchez et al. (2021) applied NLP to uncover cybercrimes on social media. Longhi (2021; 2022) extended these methods to terrorism-related discourse, demonstrating the forensic utility of computational approaches. However, these systems share a common architectural limitation: They assume that criminal intent is lexically explicit. As Gambetta (1993, 2009) demonstrates, organized-crime communication operates on precisely the opposite principle: vagueness signals loyalty, and explicit threat markers are absent by design. This architectural mismatch produces systematic false negatives on coded criminal discourse.

Forensic Linguistics, which examines language as evidence, has extensively documented the evidentiary challenges of criminal discourse. Shuy's foundational work (1993; 2005; 2014) established frameworks for analyzing language crimes, courtroom evidence, and murder cases from a linguistic perspective. Fraser (2003; 2022) and Richardson et al. (2022) have examined transcription reliability, a critical concern for any pipeline operating on wiretap data. Tiersma & Solan (2012) analyzed the language of crime from a legal-linguistic interface, while Guillén-Nieto & Stein (2022) provided a comprehensive framework for language as evidence. Yet Forensic Linguistics has primarily treated transcripts as objects to be authenticated rather than as corpora to be computationally analyzed for ideological structure.

Research on Mafia discourse has remained predominantly qualitative within Linguistics. Mannino et al. (2015) examined communication strategies in Cosa Nostra through empirical interviews; Poppi & Di Piazza (2017) analyzed discursive strategies and ideology using CDA; Trumper et al. (2014) documented lexical codes across Italian criminal organizations; Paternostro (2018) studied Mafia language in its spoken and unspoken dimensions. These studies confirm the systematic nature of criminal discourse but lack the computational operationalization required for automated detection. Morandini (2025) proposed initial NLP tools for Mafia language analysis, providing the methodological foundation for the pipeline this paper extends.

The present pipeline bridges these traditions by formalizing van Dijk's (2011) CDA framework as a sequential NLP architecture, enabling automated detection of discourse whose criminal force derives from organizational context rather than form.

3 Theoretical Grounding

Four frameworks converge in the pipeline: anti-language explains why criminal discourse is coded; socio-cognitive CDA provides the macro–meso–micro analytical structure; Speech Act Theory captures intent beyond syntax; and Costly Signal Theory explains why context-free classifiers fail.

Anti-language describes a coded linguistic variety consciously generated by what Halliday (1976) terms an anti-society: a subculture that exists in opposition to mainstream society. Its primary functions are to maintain organizational secrecy, establish ingroup solidarity, and render communication opaque to outsiders. The principal mechanism is relexicalization: standard terms are systematically replaced with coded equivalents. In American Mafia wiretap transcripts, for instance, "package" denotes contraband and "straighten out" denotes murder. Because surface syntax remains grammatically standard, traditional threat-detection models fail: criminal intent is encoded through context-dependent implicature rather than explicit markers.

To computationally map these implicit meanings, the pipeline employs van Dijk's socio-cognitive approach to CDA. The anti-language is not a direct reflection of society, but is mediated by a cognitive interface of mental models, group beliefs, and social representations. A central mechanism of this framework is the Ideological Square, a macro-discursive strategy governed by a dual polarization: emphasize Positive Things about Us (the ingroup), emphasize Negative Things about Them (the outgroup). In organized crime structures, this polarization is highly pronounced. The "Us" (Mafia) is framed by linguistic structures that reinforce strict hierarchical loyalty. In contrast, the "Them" (rivals or LEAs) is consistently framed by discourse marked by hostility, threat, and marginalization. The 7-step decision pipeline in Appendix B translates this socio-cognitive polarization into concrete NLP subtasks, specifically by measuring features such as negativity amplification and deontic saturation (the density of commands and obligations) to flag these ideological boundaries. At the micro-level of transcript utterances, the pipeline relies on Speech Act Theory (Austin, 1962; Searle, 1969) to capture intent rather than literal syntax.

Diego Gambetta's (2009, pp. 3—29) Costly Signal Theory explains the forensic consequences of such masking: The true meaning of a communicative signal is strictly determined by the surrounding criminal organization. Consequently, an NLP classifier that lacks an overriding organizational context model will inevitably produce false negatives. These errors do not stem from a lack of training data but from an architectural mismatch in which the system evaluates sentences at the wrong analytical layer. The pipeline proposed here resolves this vulnerability by computing the macro-organizational context first.

4 Corpus and data collection

The pipeline was developed on a 14,072-word corpus extracted from 18 RICO Act court documents (2006–2023), retrieved via PACER (Public Access to Court Electronic Records) from the Districts of New York and Pennsylvania. The corpus comprises wiretapped conversations involving members of the five New York and Philadelphia families (Bonanno, Colombo, Gambino, Genovese, Lucchese).

Documents were processed using ABBYY FineReader for OCR and manually verified. Original FBI transcription conventions were preserved: [U/I] for unintelligible segments, [ph] for phonetic transcriptions, [inaudible] for background noise, and [NAME] for clarified deictic references. No normalization was applied to dialectal forms (gonna, gotta, youse), ensuring authentic micro-linguistic data.

4.1 Annotation Protocol

A hybrid annotation protocol combined computational extraction with manual functional coding. Annotation was performed by the first author in two passes, separated by 12 months, to ensure intra-annotator consistency. First, corpus tools (AntConc, Sketch Engine, spaCy) extracted high-frequency features: modal verbs (must, gotta, have to), pronominal clusters (I/we bigrams), and lexical nodes (financial terms, outgroup labels). Extracted instances were then manually coded in CATMA (Computer-Assisted Text Markup and Analysis) for pragmatic function rather than grammatical form alone. Deontic modals were tagged as Subjective Authority (speaker-imposed) vs. Objective Necessity (system-imposed). Pronouns were coded based on the semantic field of the predicate (e.g., I + Command vs. we + Violence). Outgroup terms were coded for sentiment (Neutral vs. Hostile) to calculate the Negativity Score, used in Ideological Square to assess discursive polarization between positive self-representation (Us) and negative other-representation (Them). This two-step protocol captured the deep structure of anti-language that automated tools alone would miss. For example, "This guy is no good anymore" was computationally retrieved using the keyword "guy," but manually tagged as an Implied Directive (an order to kill) rather than a simple description: a distinction critical for forensic interpretation. Genre boundaries (Initiation Ritual, Sit-Down, Coded Introduction) were annotated at the document level following the move structures validated in Morandini (2026).

5 The Detection Pipeline

The seven-step decision tree in Annex A was developed as a replicable analytical procedure for human analysts (Morandini, 2026). Its structure is explicitly sequential: Each step produces an OUTPUT that constrains the interpretation space

available to the next. This sequential constraint logic is what distinguishes it from a flat feature-extraction approach and makes it a natural candidate for a pipeline architecture. Table 1 maps each step to its NLP counterpart.

Corpus annotation, feature extraction tools (spaCy, NLTK, Sketch Engine), statistical validation (χ^2 tests), and a custom relexicalization detector are fully implemented and validated on the pilot corpus. The machine learning classifiers (fine-tuned RoBERTa (Liu et al., 2019), CRF sequence labeler, domain-adapted NER) and the integrated ALI scoring system are proposed architectures requiring training data beyond the current 14,072-word corpus.

#	Level	Analytical Task	NLP Subtask	Key Output
1–2	Macro	Context model; ideological polarization; ingroup/outgroup actors	NER + speaker-role classification; sentiment polarization	Actor graph; ingroup/outgroup labels
3	Meso—Social Practices	Social practice identification (Alternative Governance, Alternative Justice, Jurisdictional Autonomy)	Multi-label doc classification (fine-tuned RoBERTa)	Practice label per conversation segment
4	Meso—Communication Genres	Genre detection: Initiation Ritual, Sit-Down, Coded Introduction	Sequence labeling (CRF/BERT) on move structure	Genre label + move boundary markers
5–6	Micro—Pragmatic and Linguistic Cues	Discursive strategies: pragmatic, syntactic, lexicosemantic cues	Rule-based + ML feature extraction (spaCy/NLTK)	6-feature vector (see Table 2)
7	Synthesis	Triangulation; anti-language probability score	Weighted ensemble scoring	ALI[1] per utterance

Table 1: Decision tree mapped to NLP subtasks.

5.1 Steps 1–2 Macro: Context and Polarization

Steps 1–2 establish the context model: the communicative event, the actor network, the power relations, and the ingroup/outgroup polarization. Computationally, this requires two sub-tasks. Named Entity Recognition (NER) with a domain-adapted model (spaCy fine-tuned on RICO annotations) identifies person references, organizational roles (boss, capo, soldier), and family affiliations. Speaker-role classification assigns hierarchical positions based on turn-taking patterns and pronominal agency. The 4.8:1 I:we ratio, indicating a high level of individual authority assertion, characteristic of hierarchical criminal organizations, was validated in the corpus and provides a quantitative precedent for authority assignment (Morandini, 2026). Sentiment polarization using Sketch Engine keyness against the OANC (Open American National Corpus) Spoken reference corpus generates ingroup/outgroup labels and flags ideological square activation when the Negativity Amplification Ratio exceeds 1.5×, a threshold derived from van Dijk's (2011) observation that ideological polarization becomes discursively salient when outgroup negativity substantially exceeds ingroup negativity. The corpus value (2.35×) exceeds this threshold, confirming robust polarization.

5.2 Step 3 Meso: Social Practice Classification

Step 3 identifies which of the three institutionalized social practices governs interaction in an *anti-state* ideological setting: Alternative Governance, Alternative Justice, or Jurisdictional Autonomy (Morandini, 2026). Each practice instantiates a distinct information structure and vocabulary cluster. In the proposed architecture, a fine-tuned RoBERTa multi-label classifier, trained on CATMA-annotated segments, predicts the practice label for each conversational turn. Practice classification constrains genre interpretation in Step 4: a Sit-Down cannot occur within a Jurisdictional Autonomy context; a Coded Introduction cannot occur within an Alternative Governance context. OUTPUT 3: practice label per segment.

5.3 Step 4 Meso: Genre Identification

Step 4 identifies the communicative genre: the conventionalized event format that activates the

[1] The Anti-Language Index (ALI) functions as a weighted ensemble score that balances top-down qualitative constraints with bottom-up quantitative features. The weighting logic (specifically, the 0.40 weight for micro-level cues) is designed to minimize false positives by ensuring that aggressive or violent language is flagged as anti-language only when it occurs within a validated organizational context model.

context model required for utterance interpretation. Three genres were documented in the RICO corpus, each with a verified move structure detailed in Appendix B (Morandini, 2026):

The Initiation Ritual enacts Alternative Governance by transforming civilian identity into organizational membership.

The Sit-Down is the primary host for Veiled Threats: the Sanction move activates the illocutionary force of implied commissives that are forensically invisible outside the genre.

The Coded Introduction enacts Jurisdictional Autonomy by communicating organizational membership precisely by not stating it: "a friend of ours" vs. "a friend of mine" encodes the binary distinction on which criminal network access depends.

In the proposed pipeline, move structure detection would use a sequence-labeling CRF trained on CATMA move-boundary annotations. OUTPUT 4: genre label + move boundary sequence.

5.4 Steps 5–6 Micro: Feature Vector Extraction

Steps 5–6 extract the six-feature vector (F1–F6 in Table 2) that constitutes the quantitative signature of *anti-language* encoding.

5.4.1 The Ideological Square as a Multi-Dimensional Feature Space

The F1–F6 vector operationalizes van Dijk's (2011) Ideological Square not as a binary classification label but as a biased distribution across two measurement axes:

Axis A (Ingroup Favoritism): High-density activation of F1 (Positive Self-Description) + F2 (Deontic Saturation/Authority Claims).

Axis B (Outgroup Derogation): High-density activation of F5 (Violence Density) + F6 (Negativity Amplification).

The pipeline identifies an Active Ideological Square when the vector simultaneously peaks in both quadrants. Features F3 (Modal Divergence) and F4 (Relexicalization Density) function as organizational filters, distinguishing criminal anti-language from generic aggressive discourse. This architecture calculates the discursive distance between "Us" and "Them" as the primary computational indicator of criminal enterprise cohesion. Table 2 maps each feature to its NLP extraction method.

Feature	Axis	Linguistic Concept	NLP Extraction Method
F1	A	Positive Self-Description	Sentiment + Speech Act Classification
F2	A	Authority/Obligation	POS Tagging (Modals) + Dependency Parsing
F3	Filter	Organizational Boundary	Chi-Square on Modal Distributions
F4	Filter	Anti-Language Encoding	Domain-Specific WSD + TTR
F5	B	Negative Other-Description	Violence Lexicon + Context Windows
F6	B	Negativity Amplification	Keyness Analysis (Log-Likelihood)

Table 2: Vector Logic—Linguistic features mapped to NLP extraction methods.

While Table 2 specifies the extraction methodology, Table 3 reports the corpus values obtained using these methods. The thresholds in Table 3 were derived empirically from the RICO pilot corpus and represent exploratory benchmarks rather than a priori parameters. The ALI weights reflect van Dijk's (2011) analytical hierarchy: Macro context and Meso social practices/communication genres are necessary but not sufficient conditions; the ideological square features (F5–F6) carry the heaviest weight (0.40) because they constitute the primary linguistic signature distinguishing criminal anti-language from aggressive-but-legitimate discourse. The ≥0.70 threshold was calibrated against manually annotated utterances from prosecuted RICO cases, selecting the value that maximized precision, flagging only utterances accepted as evidence, while maintaining recall above 0.80.

Feature	Metric	Corpus Value	*Anti-Language* Index
F1: C:E ratio	Commissive: Expressive	1.25:1	Positive if > 1.0—enforced reciprocity
F2: Deontic saturation	Deontic:Epistemic ratio	3.8:1	Positive if > 2.0—grammar of obligation
F3: Modal div.	χ^2 intra/inter-family	18.42	Positive if sig.—organizational filter active
F4: Relexicalization density	Impact: Business ratio	13.3:1	Positive if > 5.0—anti-language encoding
F5: Violence density	Violent terms/100 w.	4.6	Positive if > 2.0—enforcement context
F6: Neg. amplification	Outgroup/In group ratio	2.35×	Positive if > 1.5—ideological square active

Table 3: Six-feature micro-level vector (Steps 5–6).

Extraction tools: NLTK and spaCy for deontic modal tagging (F2), pronominal agency parsing

(F1), and violence-term lookup (F5). Sketch Engine log-likelihood computes negativity amplification (F6). A custom Python relexicalization detector addresses F4 using domain-specific substitution tables (see Appendix C). Chi-square testing on turn-level modal distributions computes F3 across family boundaries. OUTPUT 5 + 6: six-dimensional feature vector per utterance.

5.5 Step 7 Synthesis: Anti-Language Index

In the proposed pipeline, the ALI aggregates all prior outputs as a weighted ensemble score (Table 3); only utterances scoring ALI ≥ 0.70 are flagged for analyst review.

6 Preliminary Evaluation

The pipeline's feature extraction components were validated against the manually annotated RICO corpus. This section reports corpus-level metrics that confirm the six-feature vector (F1–F6) captures statistically significant patterns consistent with anti-language characteristics.

6.1 Feature Vector Validation

All six features exceeded their respective thresholds, indicating that the corpus exhibits the quantitative signature of anti-language encoding. The 4:1 Deontic Saturation[2] ratio (F2) reflects a "grammar of absolute obligation" where speakers frame actions as non-negotiable necessities rather than possibilities. The 13.3:1 Relexicalization Density (F4) confirms complete semantic displacement: 93% of financial collocations occur with violence vocabulary (*blast*, *straighten out*, *crack*) rather than legitimate business terms (*negotiate*, *contract*, *invoice*).

6.2 Pragmatic Divergence Test

To verify that the pragmatic patterns identified by the pipeline diverge significantly from standard professional discourse, a Chi-Square test was conducted to compare directive mitigation rates between the RICO corpus and the Santa Barbara Corpus of Spoken American English (SBCSAE). Of 149 directives extracted from the RICO corpus, 146 (98.0%) were bald-on-record—issued without hedging, politeness markers, or face-saving strategies. The three mitigated instances (2.0%) involved *please* followed by bare imperatives, functioning as surface politeness tokens rather than genuine mitigation. This contrasts sharply with SBCSAE workplace norms, where mitigation rates reach 88.6%. Pearson's Chi-Square analysis (χ^2 = 184.5, $p < .001$) confirms that the difference is statistically significant.

6.3 Cross-Family Consistency

The Modal Divergence feature (F3) tests whether linguistic patterns are organizationally determined rather than idiolectal. Chi-Square analysis of modal verb distributions across family boundaries (χ^2 = 18.42, $p < 0.001$) reveals a significant organizational filter: speakers deploy prospective threats ("I'll put you in the ground") within their own Family hierarchy but shift to retrospective forms ("I should've punched his head in") in inter-family contexts. This finding suggests that organizational membership constrains available speech act variants before individual strategy operates, precisely the context dependency that the pipeline is designed to detect.

6.4 Keyness Analysis

Log-Likelihood keyness analysis against the OANC Spoken reference corpus identified the top-ranked keywords in the RICO corpus. Critically, ideology-encoding keywords (n=283) exceeded generic taboo vocabulary (n=241), yielding a 1.17:1 ratio. The most theoretically significant keywords encode the organizational ideology identified at the macro level, as in *gotta* (LL=2,590.1; deontic modality), *skipper* (hierarchy), *forgiveness* (LL=276.4; honor rituals), and *straighten out* (LL=90.7; violence euphemism). This distribution distinguishes the corpus as a distinct anti-language rather than merely informal criminal speech, validating the relexicalization detector (F4) as a discriminating feature.

6.5 Pronominal Agency Shift

The corpus exhibits a 4.8:1 *I*:*we* ratio (673:141), indicating patriarchal concentration of agency. This finding empirically contradicts the organization's

[2] Deontic Saturation measures the density of the "grammar of obligation" (e.g., must, shall, have to). In an NLP context, this is extracted via dependency parsing and Part-of-Speech (POS) tagging to identify modal verbs that function as indicators of hierarchical power and non-negotiable directives, which are key signatures of organized crime authority nodes.

own mythology of "Brotherhood": authority is linguistically concentrated in singular agents, as in bosses issuing directives, while collective pronouns appear primarily in obligation contexts, as in "we gotta do this". This asymmetry provides the quantitative prior for speaker-role classification in Steps 1–2: speakers with high *I*-authority ratios are classified as hierarchical superiors; speakers with high obligatory-*we* frequency are classified as subordinates.

7 Discussion and Limitations

The pipeline's primary limitation is corpus size. At 14,072 words from 18 transcripts, the RICO pilot corpus is scarce for training deep learning components at Steps 3–4 without transfer learning from larger criminal discourse datasets. A related limitation concerns threshold derivation: feature thresholds (including the 1.5× Negativity Amplification Ratio) were established post hoc from this corpus, creating circularity in corpus-based validation. The 1.5× threshold was selected as the minimum ratio at which outgroup negativity becomes statistically distinguishable from baseline; the observed value (2.35×) substantially exceeds this floor. Future work will validate thresholds on held-out data from additional RICO transcripts.

The χ^2 tests compare RICO against external corpora (OANC, SBCSAE), providing independent evidence of distinctiveness, but thresholds require validation on held-out data. Expanding the annotated corpus across families, decades, and digital media types is the immediate priority.

Ground-truth validation presents a second challenge specific to this domain. RICO conviction records provide one verification layer; transcripts from successfully prosecuted cases confirm that flagged utterances were accepted as criminal evidence. However, negative cases (intercepted communications from organizational members that do not result in prosecution) are essential for false-positive calibration and are not available through PACER. Cross-institutional collaboration with LEAs is required to access this material in accordance with appropriate legal protocols.

Evidentiary admissibility requires that pipeline outputs be interpretable as expert evidence rather than black-box scores. The pipeline's step-by-step architecture addresses this directly: Each ALI component can be narrated as an analytical reasoning chain (Solan, 2013): this utterance scores 0.82 because the context model is active (Steps 1–2), it occurs in the Sanction move of a Sit-Down (Step 4), and F1, F2, and F4 are all above threshold, rather than requiring the court to accept an opaque numeric output.

The digital extension of the pipeline is both urgent and tractable. The relexicalization system is designed to meet the Mafia's code of silence, known as *Omertà*, which imposes strict rules around information sharing in written digital communications. This specialized language acts like its own form of encryption. The same set of features (F1–F6) used for encrypted messaging also applies here; we just need to adapt the Named Entity Recognition (NER) and relexicalization components to fit the digital context.

8 Three LEA Applications

This paper builds upon three core operational tasks identified during the longitudinal validation of the RICO corpus.

(1) Automated flagging. High-volume interception operations generate more transcript data than human analysts can review. The pipeline provides triage: ALI scores above threshold queue transcripts for priority review, reducing analyst burden while preserving human interpretive judgment at the evidentiary stage.

(2) Context-sensitive decoding. The relexicalization detector (F4) provides automatic substitution table lookup: "work" in an enforcement-context passage, identified by $F5 > 2.0$, is flagged with its decoded equivalent. This operationalizes the forensic pragmatics finding that surface assertives encode deep directives and commissives (Morandini, 2026), precisely the evidentiary gap in RICO proceedings that "reasonable person" interpretation fails to bridge.

(3) Communication network analysis. Mapping ALI scores and role labels (Steps 1–2) across speakers and transcripts reconstructs the organizational hierarchy from discourse patterns rather than informant testimony. Speakers with consistently high deontic-saturation (F2) and low mitigation rates are authority nodes; speakers with high obligatory-*we* frequency are subordinate nodes. This network topology is independently verifiable against RICO indictment records.

9 Future Work

Several research directions emerge from this initial specification. The immediate priority is expanding the annotated corpus across families, decades, and jurisdictions. Incorporating RICO transcripts from Chicago, Detroit, and New England families would test the generalizability of the feature thresholds derived from the New York/Philadelphia pilot corpus.

As Mafia communication migrates to encrypted messaging platforms (Signal, WhatsApp, Telegram), the pipeline requires domain adaptation for digital written discourse. Preliminary analysis suggests that *Omertà*'s informational constraints, which produce similar anti-language patterns in text-based communication, are also present in seized device data. The feature vector (F1–F6) architecture is medium-agnostic; adaptation requires retraining the NER component on digital naming conventions and updating the relexicalization detector for platform-specific vocabulary.

The analytical framework developed here, integrating CDA, Speech Act Theory, and corpus-based methods, offers a replicable model for analyzing other criminal organizations. Applying the same methodology to Italian mafias ('Ndrangheta, Camorra, Sacra Corona Unita), as well as to international groups (Yakuza, Russian *Organizatsiya*, Chinese Triads), would test the generalizability of anti-language theory across organizational types and linguistic systems. A cross-linguistic analysis of Sicilian, Italian, and American Mafia discourse may illuminate how anti-language adapts across linguistic environments while maintaining ideological coherence.

10 Conclusions

This paper has proposed a context-first NLP detection pipeline for automated anti-language identification in RICO transcripts and encrypted criminal communications. By formalizing van Dijk's (2011) seven-step decision tree as a sequential NLP architecture, from speaker-role classification through genre detection to ensemble scoring, the pipeline resolves the architectural mismatch that causes current classifiers to fail on coded criminal discourse: it establishes organizational context before classifying utterances, replicating the inferential logic that criminal illocutionary force requires. The six-feature micro-level vector (F1–F6), empirically grounded in a validated RICO corpus, provides the quantitative operationalization of the ideological square. The pipeline's step-by-step transparency makes its outputs admissible as expert evidence, a requirement no existing black-box classifier meets for forensic NLP applications.

Acknowledgments

This research is part of the author's doctoral dissertation at the Universidad de Alicante, supervised by Professor Victoria Loreto Guillén-Nieto. The author thanks the anonymous reviewers for their constructive feedback.

References

John Langshaw Austin. 1962. How to Do Things with Words. Oxford University Press, Oxford.

John W. Du Bois, Wallace L. Chafe, Charles Meyer, Sandra A. Thompson, Robert Englebretson, and Nii Martey. 2000–2005. Santa Barbara Corpus of Spoken American English, Parts 1–4. Linguistic Data Consortium, Philadelphia. https://doi.org/10.21415/T5VG6X

Helen Fraser. 2003. Issues in transcription: Factors affecting the reliability of transcripts as evidence in legal cases. Speech, Language and the Law, 10(2):203–226. https://doi.org/10.1558/sll.2003.10.2.203

Helen Fraser. 2022. A Framework for Deciding How to Create and Evaluate Transcripts for Forensic and Other Purposes. Frontiers in Communication, 7:1–14. https://doi.org/10.3389/fcomm.2022.898410

Evelyn Gius, Jan Christoph Meister, Marco Petris, Christian Bruck, Janina Jacke, Mareike Schumacher, Dominik Gerstorfer, Marie Flüh, and Jan Horstmann. 2023. CATMA 7.0. Zenodo. https://doi.org/10.5281/zenodo.1470118

Diego Gambetta. 1993. The Sicilian Mafia: The Business of Private Protection. Harvard University Press, Cambridge, MA.

Diego Gambetta. 2009. Codes of the Underworld: How Criminals Communicate. Princeton University Press, Princeton, NJ. https://doi.org/10.1515/9781400833610

Janet Giltrow and Dieter Stein. 2009. Genres on the Internet: Innovation, evolution, and genre theory. In Giltrow & Stein (eds.), Genres in the Internet, pp. 1–26. John Benjamins, Amsterdam. https://doi.org/10.1075/pbns.188.

Victoria Guillén-Nieto and Dieter Stein. 2022. Language as Evidence: Doing Forensic Linguistics. Palgrave Macmillan, London. https://doi.org/10.1007/978-3-030-84330-4

Herbert Paul Grice. 1975. Logic and conversation. In P. Cole and J. Morgan (eds.), *Syntax and Semantics 3: Speech Acts*, pp. 41–58. Academic Press, New York. https://doi.org/10.1163/9789004368811_003.

Michael Alexander Kirkwood Halliday. 1976. Anti-languages. American Anthropologist, 78(3):570–584. https://doi.org/10.1525/aa.1976.78.3.02a00050.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/arXiv.1907.11692

Adam Kilgarriff, Vit Baisa, Jan Bušta, Miloš Jakšič, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten years on. Lexicography, 1(1):7–36. https://doi.org/10.1007/s40607-014-0009-9.

Julien Longhi. 2021. Using Digital Humanities and Linguistics to Help with Terrorism. Forensic Science International. http://doi.org/10.1016/j.forsciint.2020.110564

Julien Longhi. 2022. Linguistic Approaches to the Analysis of Online Terrorist Threats. In V. Guillén-Nieto and D. Stein (Eds.), Language as Evidence, pp. 439–459. Palgrave Macmillan, London. https://doi.org/10.1007/978-3-030-84330-4_13

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. PLoS ONE, 14(8):1–16. https://doi.org/10.1371/journal.pone.0221152

Giovanni Mannino, Fabio Lo Verso, Simona Di Blasi, and Eugenio Baldari. 2015. Communication strategies used in the Cosa Nostra: An empirical study. Journal of Forensic Sciences, 60(4):1040–1044. https://doi.org/10.1080/02604027.2015.1113770

Fernando Miró-Llinares, José R. Moneva, and Miriam Esteve. 2018. Hate is in the Air! But Where? Introducing an Algorithm to Detect Hate Speech in Digital Microenvironments. Crime Science, 7(15):1–12. https://doi.org/10.1186/s40163-018-0089-1

Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, & Henning Peters. 2023. explosion/spaCy: v3.7.2: Fixes for APIs and requirements (v3.7.2). Zenodo. https://doi.org/10.5281/zenodo.10009823

Elena Morandini. 2025. Mafia Language Analysis and Detection Using Natural Language Processing Tools. In V. Guillén-Nieto and D. Stein (eds.), Manual of Romance Forensic Linguistics, pp. 171–202. De Gruyter, Berlin. https://doi.org/10.1515/9783110798173-008

Elena Morandini. 2026. The language of the American Mafia: A mixed-methods approach. Doctoral Thesis, Universidad de Alicante.

Giuseppe Paternostro. 2018. La lingua della mafia nel parlato e nel taciuto. In A. De Benedetto (Ed.), Comunicare oggi, pp. 245–270. Aracne, Rome.

Franca Poppi and Salvatore Di Piazza. 2017. Discursive strategies and ideology in Mafia discourse. Journal of Language Aggression and Conflict, 5(1):1–28. https://doi.org/http://doi.org/10.1016/j.dcm.2016.11.001

José Ramírez Sánchez, Alfredo Campo-Archbold, Andrea Rozo, Daniel Díaz-López, Javier Pastor-Galindo, Félix Gómez Mármol, and Jesús A. Díaz. 2021. Uncovering Cybercrimes in Social Media through Natural Language Processing. Hindawi, 2021:1–15. https://doi.org/10.1155/2021/7955637

Emma Richardson, Kate Haworth, and Felicity Deamer. 2022. For the Record: Questioning Transcription Processes in Legal Contexts. Applied Linguistics, 43:677–697. https://doi.org/10.1093/applin/amac005

John Searle. 1969. Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press, Cambridge.

Roger Wellington Shuy. 1993. Language Crimes: The Use and Abuse of Language Evidence in the Courtroom. Blackwell, Oxford.

Roger Wellington Shuy. 2005. Creating Language Crimes: How Law Enforcement Uses (and Misuses) Language. Oxford University Press, Oxford.

Roger Wellington Shuy. 2014. The Language of Murder Cases: Intentionality, Predisposition, and Voluntariness. Oxford University Press, Oxford.

Lawrence Solan. 2013. Intuitions versus algorithms: The case of forensic authorship attribution. Journal of Law and Policy, 21(2):551–576. https://brooklynworks.brooklaw.edu/jlp/vol21/iss2/13

Peter Tiersma and Lawrence Solan (Eds.). 2012. The Oxford Handbook of Language and Law. Oxford University Press, Oxford. https://doi.org/10.1093/oxfordhb/9780199572120.001.0001

John Trumper, Antonio Nicasio, Marta Maddalon, and Nicola Gratteri. 2014. Male Lingue: Vecchi e nuovi codici della mafia. Pellegrini Editore, Cosenza.

Teun Adrianus van Dijk. 1998. Ideology: A Multidisciplinary Approach. Sage, London. https://doi.org/10.1002/9781405198431.wbeal1247

Teun Adrianus van Dijk. 2008. Discourse and Context: A Sociocognitive Approach. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511481499

Teun Adrianus van Dijk. 2011. Discourse and Ideology. In T.A. van Dijk (ed.), Discourse Studies: A Multidisciplinary Introduction, pp. 379–407. Sage, London. https://doi.org/10.4135/9781446289068.n18

Teun Adrianus van Dijk. 2015. Critical Discourse Analysis. In D. Tannen, H. Hamilton, and D. Schiffrin (eds.), The Handbook of Discourse Analysis, 2nd ed., pp. 466–485. Wiley-Blackwell, Oxford. https://doi.org/10.1002/9780470753460.ch19

Appendix A: The Pipeline

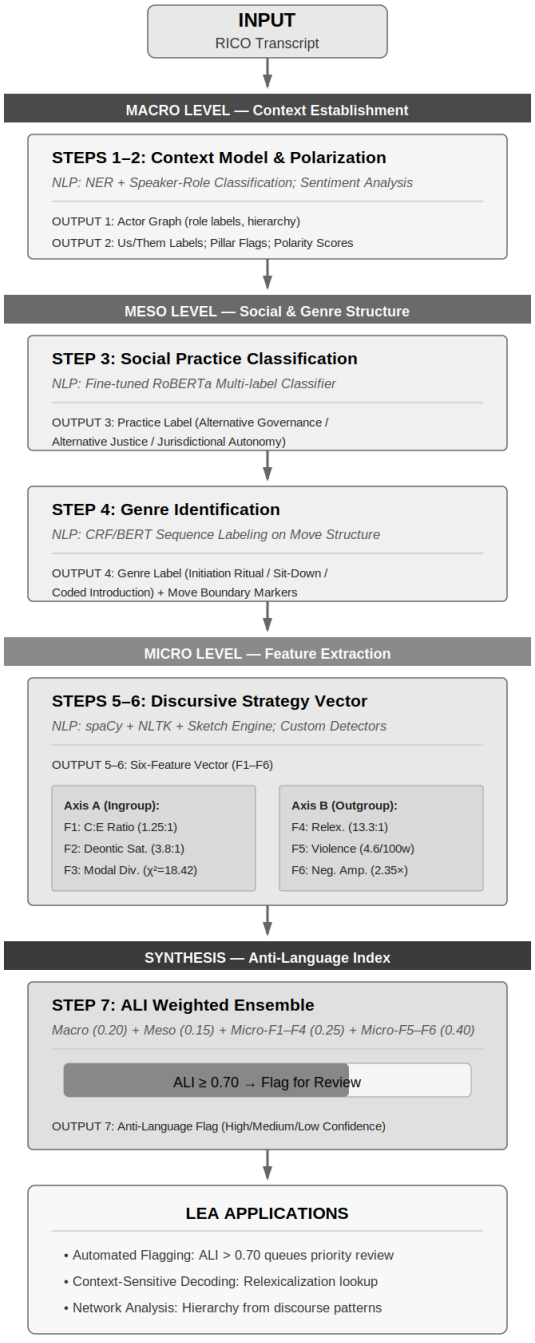

Figure A.1: Schematic architecture of the 7-step context-first Anti-Language Identification pipeline. Macro-level context (Steps 1–2) constrains Meso-level interpretation (Steps 3–4), which in turn activates Micro-level feature extraction (Steps 5–6). The ALI weighted ensemble (Step 7) aggregates all outputs; utterances scoring ≥ 0.70 are flagged for analyst review.

Appendix B: Genre Move

Genre	Move	Move Structure
Initiation Ritual	7	M1: Candidate Presentation → M2: Sponsor Vouching → M3: Boss Authorization → M4: Oath Administration → M5: Blood Pact → M6: Rule Recitation → M7: Brotherhood Confirmation
Sit-Down	7	M1: Convocation → M2: Opening Grievance → M3: Response → M4: Witness Testimony → M5: Deliberation → M6: Judgment → M7: Acceptance
Coded Introduction	4	M1: Encounter → M2: Third-party Identification → M3: Formula Exchange → M4: Access Confirmation

Table A.1: Genre move structures validated on RICO corpus (Morandini, 2026).

Appendix B: Relexicalization Table

Surface Term	Criminal Intent	n	Corpus Evidence
straighten out	murder / correct	7	straighten him out, believe me
work	homicide / assignment	12	do a piece of work
thing	weapon / conspiracy	23	that thing we discussed
friend of ours	made member	8	Coded Introduction formula
friend of mine	associate (non-member)	4	Coded Introduction formula
take care of	murder / bribe	9	take care of that problem
whack	murder	6	whack this motherfucker
blast	assault / murder	4	blast across his chest
crack	assault	3	crack his head
juice / vig	illegal interest	5	loan sharking context
earn	generate illegal income	11	He's a good earner
package	contraband / money	6	financial transactions
Violence/Impact terms (sample)		**93**	93.0% of resolution terminology
Business/Legal terms	*settle, invoice, contract*	**7**	7.0% of resolution terminology

Table A.2: Sample relexicalization mappings with corpus frequencies (n). Displacement ratio: 13.3:1 (Impact:Business). Violence density in financial dispute passages: 4.6 violent terms per 100 words.