CRIME: The Corpus of Recorded Investigative, Media, and Evidence-based Proceedings

Steven Coats[1], Dana Roemling[2]
[1]University of Oulu, Finland [2]University of Birmingham, UK
E-mail: steven.coats@oulu.fi, danaroemling@gmail.com

Abstract

This paper presents CRIME: The Corpus of Recorded Investigative, Media, and Evidence-based proceedings, a structured, searchable resource comprising audio and ASR-generated transcripts from investigative interviews, courtroom interactions, and related media. Collected from publicly available YouTube sources under the EU Data Mining Act, the corpus addresses a critical gap in current research: the lack of large-scale, real-world datasets that integrate reliable transcripts with corresponding audio. Previous studies often rely on limited data, constraining generalizability and hindering methodological innovation. By enabling detailed analysis of linguistic, phonetic, pragmatic, and discourse-level features, CRIME supports interdisciplinary research in linguistics, law, psychology, and computational modeling. Future applications include the identification of language patterns associated with interviewing strategies and outcomes, as well as leveraging large language models to explore affective and interactional dynamics. This resource offers substantial potential to inform both academic inquiry and evidence-based practices in investigative interviewing and broader criminal justice contexts.

Keywords: corpus linguistics**,** YouTube**,** forensic linguistics, investigative interviewing, large-scale discourse analysis

1. Introduction

This paper introduces CRIME – the *Corpus of Recorded Investigative, Media, and Evidence-based proceedings* – a new structured and searchable language resource designed to support research at the intersection of language, crime, and justice. CRIME brings together high-quality Automatic Speech Recognition (ASR) transcripts and audio from three distinct but related domains: police investigative interviews, courtroom proceedings, and criminal-justice-related media content. The corpus is intended to facilitate linguistic, forensic, and interdisciplinary analysis by providing access to naturally occurring spoken data across a range of legal and quasi-legal contexts. In addition to its value for forensic and legal linguistic inquiry, CRIME also contributes to the study of Computer-Mediated Communication (CMC) by enabling analysis of spoken interactions captured, processed, and transmitted through digital technologies. In what follows, we review related work, describe the design and construction of CRIME, outline the processes used to collect and curate the data, and present a sample analysis to demonstrate the corpus's potential for exploring discourse features in criminal justice settings.

2. Related Work

CRIME provides data for corpus-based research into language and discourse content in forensic linguistics, a specialized branch of applied linguistics which applies linguistic methods, approaches and knowledge to legal, investigative, and criminal contexts (see Coulthard et al., 2017). Forensic linguistics encompasses, for example, the analysis of language evidence such as ransom notes (Roemling & Grieve, 2024), but also the analysis of interview settings in legal, criminal or investigative contexts. Investigative interviewing refers to a non-coercive, evidence-based approach to interviewing suspects, witnesses, and victims, designed to gather accurate and reliable information while respecting the rights of the interviewee (see Meissner et al., 2023). Over the past decade, interest in investigative interviewing has grown significantly, marking a clear departure from more confrontational or accusatory interrogation styles (e.g., Yuan, 2010), which have been shown to increase the risk of false confessions and unreliable testimony. The field has become increasingly interdisciplinary, drawing on insights from, for example, psychology, linguistics, and policing research (Denault & Talwar, 2023), and encompasses a broad range of topics. For instance, some studies have explored how different question types shape the course and outcome of interviews (see Oxburgh et al., 2010), while others have investigated how authority and interactional asymmetries are constructed within interview discourse (e.g., Madrunio & Lintao, 2024).

Researchers have also considered how institutional roles and hierarchies are enacted in broader legal contexts. For example, Rañosa-Madrunio (2014) draws on a small corpus of five interviews from the Philippines, while Tkačuková (2010) conducts a detailed single-case study of courtroom discourse in the U.S. Others have focused on how honesty, deception, or denial are constructed and negotiated in discourse: Stokoe (2010), working with a corpus of 120 UK police interviews, analyses how men deny accusations of violence, while Benneworth-Gray (2015) explores obligations of honesty using a smaller sample of three UK interviews. Carter (2014), in a single-case analysis, questions how deception is linguistically framed. Studies have also explored the role of language mediation in investigative interviews. Filipović (2008), for instance, draws from a corpus of 10,000 pages of U.S. police transcripts involving Spanish-English interactions. Additionally, researchers have turned their attention to how age (Heini, 2023; Jol & Van der Houwen, 2014) or disability (Pereira, 2024) affect communication in legal settings.

While research into investigative interviewing has progressed significantly, a key challenge remains: the scarcity of large, systematically compiled corpora that integrate both transcripts and corresponding audio. Much

of the existing work is based on limited datasets or individual case studies, which can fall short of capturing the full complexity of real-world interactions and limit their generalizability to other contexts. Furthermore, the reliability of transcriptions is often compromised, raising concerns about the accuracy and validity of subsequent analyses (Richardson et al., 2022). To address this gap, the following section introduces the *Corpus of Recorded Investigative, Media, and Evidence-based proceedings*.

3. Corpus

The corpus was created from content hosted on the YouTube platform, using a Python-based data-collection pipeline, a method increasingly common in dialectology, sociolinguistics, and other linguistic subfields (Coats, 2023). The approach relies on the stability and widespread use of common streaming protocols such as DASH (Dynamic Adaptive Streaming over HTTP; Sodagar 2011) or HLS (HTTP Live Streaming; Pantos & May 2017), which enable content including video, audio, and transcripts to be harvested.

Two YouTube channels devoted to criminal justice content comprise the majority of the corpus: *Court TV*, a U.S. television channel founded in 1991, and *Law & Crime* Network, an internet-based content provider founded in 2017. In addition to these channels, content from four YouTube playlists was harvested: *Full length criminal interrogations*, *Law and Crime interrogations* (a subset of the content from *Law & Crime*), *Interrogation raw*, and *Trial archives*. For each video in these channels and playlists, scripts collected YouTube's own ASR transcripts, any other available transcripts uploaded for the video, and the full audio file for the episode, in .wav format. Data collection was undertaken with `yt-dlp`, an open-source Python library for harvesting content from YouTube and other platforms. After removal of duplicated content, ASR transcripts were tagged for part of speech using SpaCy's `en_core_web_sm` model (Honnibal et al. 2020); word timing tags from YouTube were retained. An overview of corpus size in terms of transcripts, word tokens, and audio duration is provided in Table 1.

Channel/Playlist	# trans.	# words (auto)	# words (other)	Length (hrs.)
Court TV	3,901	16,780,779	21,911,434	1,929.43
Law and Crime	19,111	114,723,611	6,820	1,7212.55
Law and Crime interrogations	21	113,973	0	12.35
Full length criminal interrogations	72	729,283	0	103.52
Interrogation raw	80	126,190	15,478	21.54
Trial archives	85	836,533	76,260	97.45
Total	23,270	133,310,369	22,009,992	19,376.84

Table 1: Corpus summary

The corpus provides access to transcript material from YouTube for purposes of research and education according to provisions of US and EU copyright law.[1] Two versions of the corpus exist: a static, downloadable table containing links to the source audio and transcript files, and an interactive searchable online version.

3.1 Static Corpus

The static version of CRIME[2] contains, in tabular form, parsed transcripts, metadata fields automatically retrieved for the corresponding YouTube video by `yt-dlp`, as well as links to downloadable audio. Metadata fields are recorded in the columns *Playlist, Channel, ID, Title, URL, Description* (if any)*, View Count, Duration (seconds), Uploader, Uploader ID, Uploader URL, Thumbnails, Timestamp, Release Timestamp, Availability, Live Status, Channel Verified, auto_transcript, other_transcript, wav, timed_auto, timed_other, timed_auto_words,* and *timed_other_words*. The parsed, part-of-speech-tagged ASR transcripts in the *timed_auto* column are suitable for linguistic analysis; the *timed_other* column contains non-YouTube transcripts that have been uploaded for the video.

3.2 Searchable online corpus

The online version of the corpus[3] contains the parsed, PoS-tagged YouTube ASR transcripts, the audio content, as well as most of the metadata information. The web interface provides search functionality in which transcript-linked 20-second audio segments are playable in the browser as mp3 files and downloadable. The corresponding video, as provided by the YouTube platform, can be viewed in an embedded window in the search interface. The preliminary version of the online corpus is hosted on infrastructure at Finland's Centre for Scientific Computing; it comprises a customized version of BlackLab (De Does et al., 2017), implemented using OpenShift/Kubernetes container orchestration.

The online corpus permits targeted searches for transcribed utterances from specified content types, as indicated in the metadata fields. For example, Figure 1, a screenshot from

[1] The "Fair Use" provision of US copyright law (17 U.S.C. § 107) and EU Directive 2017/790 permit reproduction and use of copyrighted materials for purposes of research and education.

[2] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/MLMB6E

[3] https://forensic.corpora.li

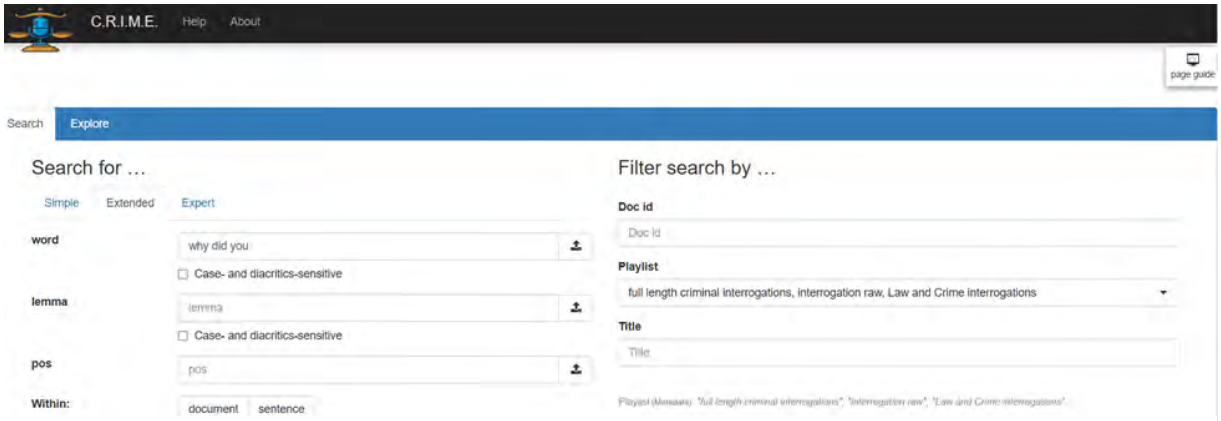

Figure 1: Extended search for *why did you* in investigative interviews

the online search interface, shows a search for the transcribed sequence *why did you* in which the "Playlist" field has been limited to the three playlists that contain transcripts of investigative interrogations.

4. Potential Analyses

To illustrate the research possibilities enabled by CRIME, this section highlights potential analytical approaches. One such area involves deontic modal and semi-modal verb forms, which express necessity or obligation and are thus essential for the effective functioning of forensic, courtroom and legal proceedings. The expression of deontic modality has undergone a shift in 20th-century English, away from the standard *must* and towards the semi-modals *have to* and *need to* (Leech, 2003; Leech et al., 2009; Mair and Leech, 2020), but the use of deontic modals and semi-modals in legal contexts has mostly been restricted to analyses of legal documents and contracts, rather than speech in investigative interviews or courtroom proceedings. CRIME offers the opportunity to investigate the use of these items in forensic speech.

Another possible application for the corpus would be to investigate the use of epistemic stance adverbials, or expressions that express and delimit a proposition's truth value in terms of semantic categories such as reality, certainty, or precision (Biber and Finegan, 1988; Hunston and Thompson, 2000). Stance markers such as *actual/actually* or *real/really* can serve to strengthen the truth value of evidential claims, while markers such as *supposedly* or *allegedly* can be used to diminish them. Attitudinal stance markers such as *honest/honestly* or *truthful/truthfully* can be used in investigative interviews to elicit specific responses as well as to strengthen claims in courtroom proceedings. While some previous studies have analyzed use of some of these items in forensic contexts (e.g. Glougie, 2016), their systematic patterning in courtroom or investigative interview discourse has mostly not been considered on the basis of evidence from larger corpora.

Figure 2 shows the relative frequencies per million transcript words of the deontic modal and semi-modals *must*, *have to*, and *need to*; the epistemic stance adverbials *actual(ly)*, *real(ly)*, *supposedly*, and *allegedly*; and the attitudinal stance markers *honest(ly)* and *truthful(ly)*, according to corpus subsection. The more formal *must* is most frequent in the *Law and Crime* channel and in the *trial archives* playlist; the former contains scripted content, while the latter comprises transcripts of court proceedings. *Have to* and *need to* are more common in investigative interviews. For the stance markers, the most striking pattern is that the attitudinal markers *honest(ly)* and *truthful(ly)* are much more common in investigative interviews, presumably as admonitions of the interviewer to the interviewee. A more in-depth study could focus on these and other frequency patterns in the corpus.

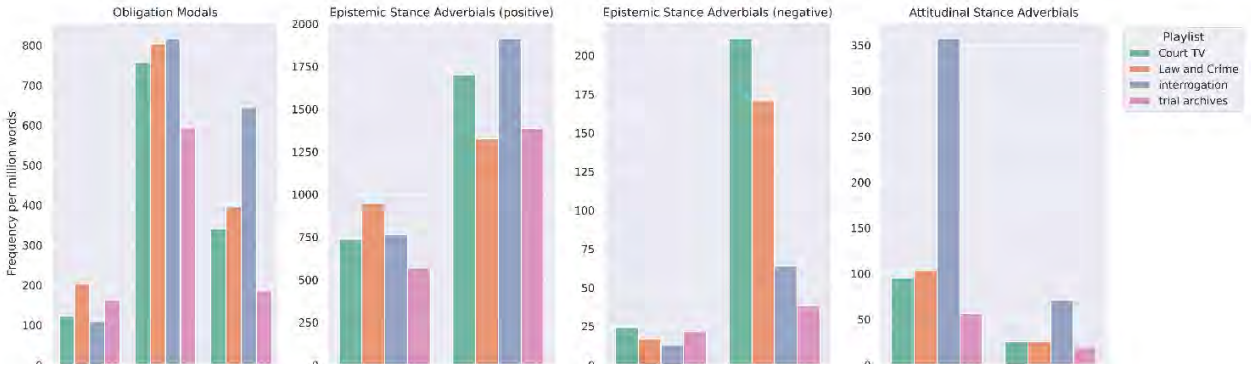

Figure 2: Relative frequencies of selected (semi-)modals and stance adverbials by corpus subsection

5. Caveats

While CRIME provides a large-scale resource for research, several caveats should be acknowledged. First, although ASR offers an efficient method for transcript generation, it is subject to transcription errors, particularly in cases of overlapping speech, strong regional accents, low audio quality, or use of out-of-vocabulary items such as some legal terminology. These errors may affect the precision of fine-grained linguistic analysis and should be taken into account when interpreting results. Second, YouTube ASR transcripts are undiarized: there are no indications of speaker turns, and therefore associating transcript segments with individual speakers requires manual annotation. Finally, as the corpus draws on publicly available YouTube channels or playlists, there is some inconsistency in the structure and content of the source material, particularly for content retrieved from the much larger *Court TV* and *Law and Crime* channels. These channels contain a mixture of content types, including "true crime" entertainment, crime and criminal justice news, commentary on court cases, and excerpts from investigative interviews and trial recordings. Researchers interested in, for example, the speech content of investigative interviews, will need to filter metadata fields such as "playlist" in order to exclude transcripts such as expert commentary or courtroom proceedings. Although some of the corpus content can be disambiguated for interaction type or for genre/register with targeted searches based on video title substrings, the variability in content type, as well as format, speaker roles, recording quality, and recording context may introduce noise or limit comparability across files.

6. Conclusion

This paper has introduced the *Corpus of Recorded Investigative, Media, and Evidence-based proceedings*. By addressing the scarcity of large-scale, real-world corpora of speech from legal contexts with aligned audio and transcript data, CRIME enables new forms of empirical analysis across disciplines like linguistics, law, psychology, and computational modeling. The corpus is already suited for a range of fine-grained linguistic investigations. For example, we highlighted how it can support the analysis of epistemic stance adverbials and deontic modal and semi-modal verb forms - features central to meaning-making in legal and investigative discourse.

Looking ahead, further development of CRIME will focus on expanding the dataset and improving transcription accuracy. New data can be added to the corpus by retrieving and parsing content that has been more recently uploaded to the targeted YouTube playlists and channels; in addition, the corpus could be expanded through the inclusion of material from other online sources. Another planned upgrade for CRIME is the implementation of larger and potentially more accurate ASR models, as well as transcript diarization, via a pipeline that incorporates Whisper, WhisperX, pyannote, or similar tools (Radford et al. 2022, Bain et al. 2023, Bredin et al. 2023). These improvements aim to enhance the corpus's value as a resource for interdisciplinary research and applied work in criminal justice communication.

7. References

Benneworth-Gray, K. (2015). 'Are you going to tell me the truth today?': Invoking obligations of honesty in police-suspect interviews. *International Journal of Speech Language and the Law*, *21*(2), Article 2. https://doi.org/10.1558/ijsll.v21i2.251

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX: Time-accurate speech transcription of long-form audio. In *Proceedings of Interspeech 2023*, pp. 4489–4493. https://doi.org/10.21437/Interspeech.2023-78

Biber, D., & Finegan, E. (1988). Adverbial stance types in English. *Discourse Processes*, 11, 1–34. https://doi.org/10.1080/01638538809544689

Bredin, H. (2023). Pyannote.audio 2.1 speaker diarization pipeline: Principle, benchmark and recipe. In *Proceedings of Interspeech 2023*, pp. 1983–1987. https://doi.org/10.21437/Interspeech.2023-105

Carter, E. (2014). When is a lie not a lie? When it's divergent: Examining lies and deceptive responses in a police interview. *Language and Law*, *1*, 19.

Coats, S. (2023). Dialect corpora from YouTube. In B. Busse, N. Dumrukcic, & I. Kleiber (Eds.), *Language and Linguistics in a Complex World* (pp. 79–102). Berlin: De Gruyter. https://doi.org/10.1515/9783111017433-005

Coulthard, M., Johnson, A., & Wright, D. (2017). *An Introduction to Forensic Linguistics: Language in Evidence* (2nd edition). Routledge, Taylor & Francis Group.

De Does, J., Niestadt, J., & Depuydt, K. (2017). Creating research environments with BlackLab. In J. Odijk & A. van Hessen (Eds.), *CLARIN in the Low Countries*. London: Ubiquity Press, pp. 245–257. https://doi.org/10.5334/bbi.20

Denault, V., & Talwar, V. (2023). From criminal interrogations to investigative interviews: A bibliometric study. *Frontiers in Psychology*, *14*, 1175856. https://doi.org/10.3389/fpsyg.2023.1175856

Filipovic, L. (2008). Language as a witness: Insights from cognitive linguistics. *International Journal of Speech Language and the Law*, *14*(2), Article 2. https://doi.org/10.1558/ijsll.2007.14.2.245

Glougie, J.R.S. (2016). *The semantics and pragmatics of English evidential expressions: The expression of evidentiality in police interviews*. Ph. D. thesis, University of British Columbia. https://doi.org/10.14288/1.0319268

Heini, A. (2023). 'Basically, I'm gonna ask you a load of questions' Cautioning exchanges in police interviews with adolescent suspects. *Language and Law 9*(2), 11–31. https://doi.org/10.21747/21833745/lanlaw/9_2a3

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *SpaCy: Industrial-strength Natural Language Processing in Python*. https://doi.org/10.5281/zenodo.1212303

Hunston, S., & Thompson, G. (2000). *Evaluation in Text:*

Authorial Stance and the Construction of Discourse. Oxford: Oxford University Press.

Jol, G. A. H., & Van der Houwen, F. (2014). Police interviews with child witnesses: Pursuing a response with maar (= Dutch but)- prefaced questions. *International Journal of Speech, Language and the Law*, *21*(1), Article 1. https://doi.org/10.1558/ijsll.v21i1.113

Leech, G. (2003). Modality on the move: The English modal auxiliaries 1961–1992. In R. Facchinetti, F. Palmer, & M. Krug (Eds.), *Modality in Contemporary English*. Berlin: De Gruyter, pp. 223–240. https://doi.org/10.1515/9783110895339.223

Leech, G., Hundt, M., Mair, C., & Smith, N. (2009). *Change in Contemporary English: A Grammatical Study*. Cambridge University Press.

Madrunio, Ma. K. J. R., & Lintao, R. B. (2024). Power, control, and resistance in Philippine and American police interview discourse. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*, *37*(2), 449–484. https://doi.org/10.1007/s11196-023-10045-8

Mair, C., & Leech, G. N. (2020). Current changes in English syntax. In B. Aarts, A. McMahon, & L. Hinrichs, (Eds.), *The Handbook of English Linguistics*. London: Wiley, pp. 249–276. https://doi.org/10.1002/9781119540618.ch14

Meissner, C. A., Kleinman, S. M., Mindthoff, A., Phillips, E. P., & Rothweiler, J. N. (2023). Investigative interviewing: A review of the literature and a model of science-based practice. In D. DeMatteo & K. C. Scherr (Eds.), *The Oxford Handbook of Psychology and Law* (1st ed., pp. 582–603). Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197649138.013.34

Oxburgh, G. E., Myklebust, T., & Grant, T. (2010). The question of question types in police interviews: A review of the literature from a psychological and linguistic perspective. *International Journal of Speech Language and the Law*, *17*(1), Article 1. https://doi.org/10.1558/ijsll.v17i1.45

Pantos, R., & May, W. (2017). HTTP Live Streaming (RFC 8216). Internet Engineering Task Force (IETF). https://doi.org/10.17487/RFC8216

Pereira, T. (2024). Establishing common ground using low technology communication aids in intermediary mediated police investigative interviews of witnesses with an intellectual disability. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*, *37*(2), 517–546. https://doi.org/10.1007/s11196-023-10035-w

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv*:2212.04356 [eess.AS]. https://doi.org/10.48550/arXiv.2212.04356

Rañosa-Madrunio, M. (2014). Power and control in Philippine courtroom discourse. *International Journal of Legal English*, *2*(1), 4–30.

Richardson, E., Haworth, K., & Deamer, F. (2022). For the record: Questioning transcription processes in legal contexts. *Applied Linguistics*, *43*(4), 677–697. https://doi.org/10.1093/applin/amac005

Roemling, D., & Grieve, J. (2024). Forensic Authorship Analysis. *CREST Security Review*, #18: Communication

Sodagar, I. (2011). The mpeg-dash standard for multimedia streaming over the internet. *IEEE Multimedia*, 18(4), 62–67.

Stokoe, E. (2010). 'I'm not gonna hit a lady': Conversation analysis, membership categorization and men's denials of violence towards women. *Discourse & Society*, *21*(1), Article 1.

Tkačuková, T. (2010). The power of questioning: A case study of courtroom interaction. *Discourse and Interaction*, 3(2), 49–61.

Yuan, C. (2010). Avoiding revictimization: Shifting from police interrogations to police interviewing in China. *International Journal of Speech Language and the Law*, *16*(2), 293–297. https://doi.org/10.1558/ijsll.v16i2.293