Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895
https://llrjournal.com/index.php/11

Closing the Gap in Corpus Linguistics: Introducing a Groundbreaking Multimodal Corpus of Informal Digital English, Encompassing Texts, Emojis, And Voice Notes

[1]Rimsha Khalid

[1]University of Sahiwal. rimsh0675@gmail.com

Liberal Journal of Language & Literature Review
Print ISSN: 3006-5887
Online ISSN: 3006-5895

Abstract

Mainstream English corpora, such as the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), predominantly focus on edited prose and formal speech. This limited approach creates a critical gap in our understanding of everyday informal digital English, particularly in the dynamic interaction between text, emojis, and short voice notes. This paper decisively addresses this gap by proposing the design of an ethically sourced, balanced, and richly annotated multimodal corpus that thoroughly captures private-by-default digital conversations, particularly those occurring in messaging apps and group chats where English is utilized. This research provides an exhaustive review of existing corpus-linguistic methodologies and establishes robust design principles for sampling, anonymization, and multimodal annotation. Furthermore, it introduces comprehensive analytical pipelines aimed at examining lexis, pragmatics, sentiment, discourse moves, and phonetic-pragmatic features in voice notes. The study precisely formulates research questions and hypotheses, presents a reproducible methodology that includes comprehensive data governance and Institutional Review Board (IRB)-ready protocols, and delineates a meticulous evaluation plan. The proposed corpus, EDDE (English Digital Discourse & Emoji), is specifically engineered to: (i) model emojis and text as co-expressive units; (ii) operationalize pragmatic functions such as stance, politeness, and mitigation; (iii) capture prosodic correlates of stance in voice notes; and (iv) provide essential insights for English as a Foreign Language (EFL) and English as a second Language (ESL) learners. This endeavor will significantly advance both fundamental and applied corpus linguistics, closing the crucial multimodality gap in the field.

Keywords: Corpus Linguistics, English digital discourse, Emoji, voice note, Text, digital corpus design, informal English.

1. Introduction

Corpus linguistics takes an assertive approach to the investigation of language through rigorous empirical analysis of large, representative samples stored digitally. The development of its tools and methods has profoundly transformed our

understanding of vocabulary and discourse. However, the current reference corpora for English predominantly emphasize edited texts and planned speech, creating a critical gap in our insights into private, informal digital interactions, such as direct messages (DMs), group chats, and mixed-media messages. This underrepresentation significantly hinders our ability to fully comprehend contemporary English. Recent discussions have effectively separated foundational corpus work encompassing data resources, tools, and methods from applied work that leverages these resources. Thus, establishing a robust multimodal corpus of informal digital English is essential for making substantial advancements in both domains.

Problem Statement: There is a significant gap in the availability of a well-rounded, accessible, and ethically shareable multimodal English corpus that effectively integrates text, emojis, and short voice notes from everyday messaging contexts under a cohesive annotation framework.

Aim: The Researcher's aim is to clearly define and robustly justify the design, governance, and analytics of EDDE—an innovative Multimodal Corpus of Informal Digital English—ensuring its implementation and evaluation as a publishable and reusable resource.

Contributions

1. A solid model for sampling and consent that prioritizes privacy in media sharing.
2. A comprehensive annotation scheme that seamlessly integrates text tokens, emojis, and prosodic features.
3. High-quality benchmarked analytic tasks, including collocation, pragmatics, sentiment analysis, and sequence analysis.
4. Impactful pedagogical applications for English for Academic Purposes (EAP) and English for Specific Purposes (ESP), along with exceptional materials design.
5. An extensive evaluation plan focusing on representativeness, inter-annotator reliability (IAR), and reproducibility, ensuring that our corpus meets the highest standards.

2. Background and Related Work

Large English corpora, such as the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), provide valuable insights into frequency, coverage, and variations in register. However, they do not adequately capture the

nuances of conversational, private digital discourse or non-textual signals. Tools like concordancers and profilers, including AntConc and AntWordProfiler, have proven essential for vocabulary research and collocation analysis. Nevertheless, support for multimodal alignment—encompassing emojis and audio—remains an area that requires further development.

The field emphasizes the importance of corpus balance, sampling, cleaning, and annotation as core design principles. This work firmly advances these principles by applying them to private digital media and setting new standards.

Gap: Existing public English datasets for social media predominantly focus on public microblogs. They overlook vital aspects such as voice notes, reduce emojis to simplistic sentiment indicators, and often encounter sharing limitations due to privacy concerns. There is a clear and pressing need for a unified, reusable corpus that effectively integrates text-emoji co-expression with short audio messages.

3. Research Questions and Hypotheses

RQ1: How do various types and positions of emojis (pre-text, intra-text, or post-text) effectively correlate with lexical and syntactic markers of stance and politeness in informal English messaging?

H1: Emojis undeniably act as powerful pragmatic operators; their co-occurrence with hedges (e.g., "maybe," "I think"), mitigators (e.g., "just," "a bit"), and modal verbs will strongly predict softer requests and expressions of positive politeness.

RQ2: What prosodic patterns in voice notes (including pitch span, intensity, and speech rate) align with textual and emoji cues of emotion and stance?

H2: Wider pitch spans and slower speech rates will consistently accompany supportive or soothing stances when enhanced by affiliative emojis (e.g., ,).

RQ3: How consistent are the sequences of discourse (opening, request, acknowledgment, closure) in dyadic versus group chats, and across various relationship types (peer, family, work)?

H3: Group chats will exhibit distinct patterns that directly reflect the dynamics of relationship types, underscoring the necessity for a nuanced understanding of digital communication.

The researcher confidently asserts that standard frequency bands and collocational profiles from formal corpora can effectively generalize to digital discourse.

H4: While high-frequency items remain prevalent, it is clear that multiword patterns and discourse markers (such as "Tysm, "Tc", "btw," and "lol") dominate local collocational spaces. Moreover, emojis serve as essential collocational hubs.

4. Corpus Design (EDDE)

4.1 Scope and Registers

Private Written Messaging: This research analysis pays particular attention to one-to-one and small group conversations (2_12 participants) as the core focus.

Voice Notes: Capped at 90 seconds, these voice notes represent dynamic, spontaneous commentary, requests, and replies, showcasing real-time interaction.

Platforms: The research utilizes opt-in exports from WhatsApp, Telegram, Signal, and Discord direct messages, ensuring absolute compliance with user terms and consent.

4.2 Sampling Frame

Strata: The researcher meticulously stratifies our sample by dyads versus groups, relationship type, time of day, and various topics such as coordination, support, work, and humor.

Targets: Researcher is set to collect an impressive 2 million word-equivalents of text, 150 hours of voice notes, and 3 million emoji tokens (including repeats) from native English speakers (L1) and proficient second-language users (L2).

Temporal: This research implements a rolling window over 12 months to capture dynamic changes and seasonal variations effectively.

4.3 Ethical Governance

- Researchers prioritize informed consent from all participants, providing granular opt-out options and a clear revocation protocol.
- Researcher processes include automated scrubbing of Personally Identifiable Information (PII), complemented by rigorous human quality assurance. We ensure synthetic replacements for names and locations where necessary.
- Researcher follows a tiered access model: open (fully anonymized text and emojis), controlled (audio features only), and restricted (raw audio stored securely).
- All regulatory documentation for Institutional Review Board (IRB) approval and Data Protection Impact Assessment is meticulously prepared and compliant.

4.4 Data Preparation

Normalization: Researcher expertly preserves original casing, elongations (e.g., "soooo"), laughter tokens (e.g., "haha," "lol"), and punctuation density to authentically represent communication.

Emoji Handling: Researcher maintains Unicode, skin-tone modifiers, and ordering, ensuring clarity by mapping to standard shortcodes while retaining position indices for accuracy.

5. Annotation Scheme

The advanced annotation scheme employs layered, stand-off annotations in JSON/XML format with integrated cross-references:

Token Layer: Researcher expertly captures words, abbreviations, hashtags, URLs, laughter, and orthographic lengthening.

Emoji Layer: Researcher precisely defines the type, cluster, position, and pragmatic functions of emojis, including acknowledgment, softening, intensifying, irony, and emotional expression.

Turn-Taking Layer: Researcher thoroughly documents speaker identifiers, timestamps, reply-to links, and thread IDs to ensure clarity and accuracy.

Discourse Move Layer: researcher framework clearly delineates various discourse moves such as requests, offers, acknowledgments, repairs, and closings.

Sentiment/Stance Layer: This research rigorously assesses valence, arousal (on a scaled basis), and stance labels (supportive, teasing, and disagreement) to provide a comprehensive analysis.

Audio-Prosody (Voice Notes): This research analyzes pitch (minimum, maximums, and means), intensity means, speech rate, and pause structures, aligning these insights with our transcript tiers.

Inter-Annotator Reliability (IAR): This Research sets a high standard for IAR, targeting Cohen's kappa $\geq .75$ for categorical tiers and the Intraclass Correlation Coefficient (ICC) for continuous tiers (prosody). Our preliminary testing involves 1,000 turns to ensure robust reliability.

6. Analytic Pipeline

Lexis & Phraseology: Researcher applies sophisticated methods, including frequency bands, type–token ratios, and Mutual Information (MI)/LogDice for text–emoji

collocations, along with impactful dispersion analyses across various strata.

Pragmatics: researchers' comprehensive inventories of hedges/mitigators, emoji roles, and sequence analyses of discourse moves reflect cutting-edge approaches that lead the field.

Prosody–Text Coupling: By employing mixed-effects models, the Researcher effectively links prosodic features to stance labels and emoji categories for deeper insights.

Register Comparison: Researcher confidently benchmark our findings against public English corpora to identify both under- and overused items and bundles, ensuring our analysis is thorough and accurate.

Reproducibility: This research's meticulous methodologies include containerized workflows, published codebooks, and sample data integrated with Jupyter notebooks, establishing robust transparency and reproducibility.

7. Methodology (Study Protocol)

7.1 Participants: Researcher will recruit between 300 and 500 adult volunteers who use English daily in digital messaging, including those who speak it as their first language (L1) as well as proficient second language (L2) speakers.

7.2 Instruments: The essential instruments will include consent forms, export guides, and comprehensive demographic and relationship questionnaires.

7.3 Procedures: Researcher will implement systematic procedures for data export, ensure robust automated personal identifiable information (PII) masking, maintain high standards of manual quality assurance, provide thorough annotation training, and execute iterative Inter-annotator Reliability (IAR) calibration.

7.4 Statistical Modeling: The Researcher will utilize Generalized Linear Mixed Models (GLMMs) for stance predictions, apply permutation tests for collocation analysis, and conduct survival analysis for response latency patterns. This approach exemplifies our commitment to rigorous research methodology.

7.5 Validation

Researcher's validation will rigorously include held-out folds by conversation and cross-register tests, which are standard and essential practices in our methodology.

8. Expected Findings and Significance

Researcher firmly asserts that emojis will systematically denote stance and politeness,

and the prosody of voice notes will show a significant correlation with affiliative stances. Furthermore, digital discourse will undoubtedly reveal unique multiword bundles and repair patterns, clearly differentiating it from formal corpora. This corpus will decisively enhance pedagogical applications, such as teaching pragmatic softeners and managing tone, while also advancing Natural Language Processing (NLP) tools for informal English by providing authoritative gold standards for emoji-text alignment.

9. Limitations

The researcher recognizes the potential for sampling bias towards users who are adept with digital platforms, as well as privacy issues that restrict the release of raw audio data. Nevertheless, the robustness of the methodologies employed effectively mitigates these challenges. This research also addresses the potential subjectivity in the annotation of pragmatic labels and acknowledges the cross-cultural variations within the English language, which are essential to consider.

10. Implementation Timeline

Months 1–2: The researcher will secure IRB approval, create datasheets, and initiate recruitment and pilot experiments without delay.

Months 3–4: The researcher will develop the annotation manual, conduct pilot inter-annotator reliability testing, and refine the annotation scheme to ensure precision.

Months 5–8: Full data collection and annotation will take place, followed by thorough interim analyses to assess progress.

Months 9–10: The researcher will concentrate on modeling and validation to ensure the integrity of the findings.

Months 11–12: Finally, the corpus will be released in a tiered format, and manuscripts will be submitted for publication, reinforcing the project's impact.

11. Conclusion

The researcher is resolutely dedicated to constructing and analyzing a balanced, ethically shareable multimodal corpus of informal digital interactions that seamlessly integrates text, emojis, and voice notes. This initiative decisively fills a critical gap in English corpus linguistics and will significantly advance the infrastructure and applied insights in vocabulary, pragmatics, pedagogy, and natural language processing (NLP). This work is perfectly aligned with contemporary trends, where multimodality plays

an intrinsic and vital role in meaning-making.

References

Anthony, L. (2017). Corpus linguistics and vocabulary: A commentary on four studies. Vocabulary Learning and Instruction.