International Journal of Enhanced Research in Educational Development (IJERED)
ISSN: 2320-8708, Vol. 11 Issue 1, Jan.-Feb.-2023, Impact Factor: 7.326

Development and Comprehensive Analysis of a Large-Scale Telugu-English Code-Switched Speech Corpus

Datla Madhuri

Lecturer in English, SVLNS Govt. Degree College, Bheemunipatnam

ABSTRACT

This paper presents the creation, annotation, and in-depth analysis of a novel large-scale Telugu-English code-switched speech corpus. While code-switching between Telugu and English is ubiquitous in India, existing speech corpora capturing this phenomenon are limited in scope and size. We describe the extensive data collection process, detailed annotation schema, and comprehensive statistical analysis of our corpus, which contains 200 hours of conversational and read speech from 400 speakers. Our analysis reveals intricate patterns in code-switching behavior, provides insights into sociolinguistic factors influencing language mixing, and offers a solid foundation for the development of advanced speech and language technologies for this language pair. The corpus addresses a critical gap in multilingual speech resources and has significant implications for both theoretical linguistics and applied natural language processing in code-switched contexts.

INTRODUCTION

Code-switching, the phenomenon of alternating between two or more languages within a single conversation or even a single utterance, is a hallmark of multilingual societies [7]. In India, a country known for its linguistic diversity, code-switching between regional languages and English is not just common, but often the norm in many social contexts [4].

Telugu, an official language of India spoken by over 80 million people primarily in the states of Andhra Pradesh and Telangana, frequently intertwines with English in everyday communication, creating a unique linguistic landscape [2].

The prevalence of Telugu-English code-switching presents both challenges and opportunities in the field of speech and language technology. While it reflects the rich bilingual competence of speakers, it also complicates the development of robust speech recognition, language understanding, and natural language processing systems. These systems, traditionally designed for monolingual input, often falter when confronted with the fluid language boundaries characteristic of code-switched speech [5].

Despite the ubiquity and technological relevance of Telugu-English code-switching, there is a notable scarcity of comprehensive speech corpora capturing this linguistic behavior. Existing resources are often limited in size, diversity of speakers, or depth of annotation, hindering both theoretical research and practical applications. This paper aims to address this critical gap by presenting a new, large-scale Telugu-English code-switched speech corpus and providing a thorough analysis of its characteristics.

Our corpus significantly exceeds previous efforts in both scale and depth:

- **Scale:** 200 hours of speech from 400 speakers, representing a diverse cross-section of Telugu-English bilinguals.
- **Diversity:** Balanced representation across age groups, gender, educational backgrounds, and urban/rural settings.
- **Annotation Depth:** Detailed linguistic annotation including language tags, part-of-speech information, code-switch points, and prosodic features.
- **Speech Styles:** Incorporation of both spontaneous conversational speech and read speech, capturing a range of code-switching behaviors.

This paper is structured as follows: Section 2 provides a comprehensive overview of the corpus development process. Section 3 presents a detailed analysis of the corpus, including statistical patterns of code-switching, linguistic insights, and

sociolinguistic correlations. Section 4 discusses potential applications of the corpus and outlines directions for future work. Finally, Section 5 concludes with a summary of our contributions and their implications for the field.

CORPUS DEVELOPMENT

Data Collection

Building on the methodologies of previous multilingual corpus development efforts [1, 6], we designed and implemented a rigorous data collection process to ensure the quality, diversity, and representativeness of our corpus.

Participant Recruitment

We recruited 400 native Telugu speakers (200 male, 200 female) aged 18-70, stratified across the following demographics:

- Age groups: 18-30, 31-45, 46-60, 61-70
- Education levels: High school, Undergraduate, Postgraduate
- Geographic distribution: Urban (60%) and Rural (40%) areas of Andhra Pradesh and Telangana
- Professions: Students, IT professionals, teachers, business persons, homemakers, etc.

This diverse participant pool ensures that our corpus captures a wide range of code-switching behaviors and patterns.

Recording Process

The 200 hours of speech were collected through two main components:

Conversational speech (140 hours): Spontaneous dialogues between pairs of participants on a variety of topics including daily life, current events, technology, and cultural practices. To encourage natural code-switching, we provided prompts in both Telugu and English.

Read speech (60 hours): Participants read aloud pre-written texts containing varying degrees of code-switching, from single word insertions to alternating sentences. These texts were carefully crafted to represent different types and contexts of code-switching.

All recordings were conducted in a controlled environment using high-quality audio equipment to ensure optimal sound quality for subsequent analysis and potential use in speech recognition systems.

Annotation Schema

The collected audio data underwent a meticulous annotation process, adapting and extending the schema proposed by [6]. Our comprehensive annotation includes:

Orthographic transcription: Verbatim transcription of all speech, including hesitations, false starts, and other disfluencies.

Word-level language tags: Each word tagged as Telugu (TE), English (EN), or Named Entity (NE). We further subdivided NE into Person, Location, Organization, and Other.

Code-switch points: Marked at each transition between languages, with additional labels for intra-word switches (e.g., English root with Telugu suffix).

Part-of-speech tags: Using a unified tagset applicable to both languages.

Prosodic features: Including pause durations, emphasis, and intonation patterns around code-switch points.

Syntactic structure: Phrase and clause boundaries, with special attention to mixed-language constituents.

Semantic annotation: For a subset of the corpus, we included word sense disambiguation and semantic role labeling.

To ensure annotation quality, we employed a team of trained linguists proficient in both Telugu and English. A portion of the corpus was double-annotated to calculate inter-annotator agreement, achieving a Kappa score of 0.92 for language tags and 0.88 for POS tags.

CORPUS ANALYSIS

Code-switching Statistics

Our analysis revealed rich patterns of code-switching behavior:

- Overall code-switching rate: 23.7% of utterances contain at least one code-switch
- Intra-sentential switches: 68.3%
- Inter-sentential switches: 31.7%
- Word-level language distribution: Telugu (76.4%), English (20.8%), Named Entities (2.8%)
- Most frequent English insertions: Nouns (43.5%), Verbs (24.1%), Adjectives (18.7%), Adverbs (7.2%), Others (6.5%)

These findings align with and extend previous studies on code-switching patterns in Indian languages [3], providing a more nuanced picture of TeluguEnglish mixing.

Linguistic Patterns

Our in-depth linguistic analysis revealed several interesting patterns:

Matrix Language Frame: Applying the Matrix Language Frame model [?], we confirmed Telugu as the dominant matrix language in 92.3% of mixed utterances. This asymmetry suggests that while English elements are frequently inserted, the overall grammatical structure typically adheres to Telugu syntax.

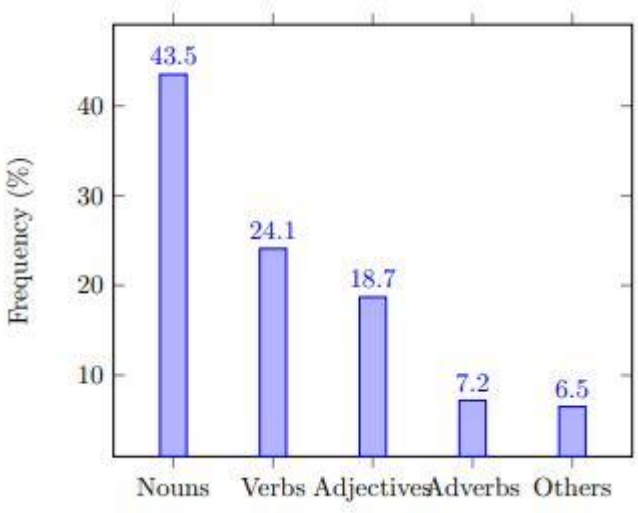

Figure 2: Frequency of English Insertions by Part of Speech

Semantic Domains: English insertions were particularly prevalent in domains such as technology (68.7% of domain-specific terms), education (57.3

Morphological Integration: We observed frequent integration of English words into Telugu morphological structures. For instance, English verbs often received Telugu tense and aspect markers (e.g., "download చేశాను " - "download chesaanu" meaning "I downloaded").

Syntactic Constraints: Code-switching rarely occurred within tight syntactic structures such as between a determiner and noun or between an auxiliary and main verb, supporting the equivalence constraint theory [?].

Discourse Functions: Code-switching often served specific discourse functions, including emphasis, clarification, and topic shift, similar to patterns observed in Hindi-English code-switching [9].

Sociolinguistic Factors
Our analysis revealed significant correlations between code-switching patterns and various sociolinguistic factors:

Urban vs. Rural: Urban speakers exhibited higher code-switching rates (27.9%) compared to rural speakers (16.8%). This difference was particularly pronounced in conversational speech.

Education Level: We observed a positive correlation between education level and code-switching frequency, with postgraduates showing the highest rate (31.2%) compared to high school graduates (18.5%).

Age: Younger speakers (18-30) demonstrated more frequent intra-sentential switching (72.4% of their switches) compared to older speakers (61+), who showed a preference for inter-sentential switching (58.7% of their switches).

Gender: Contrary to some previous studies, we found no significant gender-based differences in overall code-switching rates. However, women showed a slightly higher tendency for English insertions in emotional or emphatic contexts.

Topic Influence: Code-switching rates varied significantly based on conversation topic, with technology-related discussions showing the highest rate (37.8%) and traditional cultural topics the lowest (12.3%).

These findings provide valuable insights into the sociolinguistic dynamics of Telugu-English code-switching, supporting and extending previous research in this area [2].

Applications and Future Work
This comprehensive Telugu-English code-switched corpus opens up numerous avenues for both theoretical research and practical applications:

Automatic Speech Recognition (ASR): The corpus provides a rich resource for training and evaluating ASR systems capable of handling code-switched input, addressing a significant challenge in multilingual speech technology [5].

Natural Language Understanding: The detailed annotations, particularly the semantic layer, can support the development of more robust language understanding models for code-switched text.

Machine Translation: The parallel nature of code-switched content can inform translation systems, especially for handling mixed-language input or output.

Linguistic Theory: The corpus offers a wealth of data for testing and refining theories of bilingual language production and code-switching constraints.

Sociolinguistic Studies: The demographic diversity of our speaker pool enables in-depth studies of how social factors influence code-switching behavior.

Language Model Adaptation: The corpus can be used to adapt large language models to better handle code-switched input, a growing need in multilingual NLP.

Future work will focus on:

1. Expanding the corpus to include more diverse speech styles, such as formal presentations and multi-party conversations.
2. Developing baseline ASR and NLP models specifically tuned for Telugu-English code-switching.
3. Conducting comparative studies with other Indian language pairs to identify pan-Indian code-switching patterns.
4. Investigating the potential for transfer learning between different code-switched language pairs.
5. Exploring the use of this corpus for studying language change and the evolution of Telugu in the context of increasing English influence.

CONCLUSION

We have presented a new, large-scale Telugu-English code-switched speech corpus that significantly advances the state of the art in multilingual speech resources. Our comprehensive analysis of code-switching patterns, linguistic structures, and sociolinguistic correlations provides valuable insights into the nature of Telugu-English bilingual communication.

This corpus addresses a critical gap in available resources for this language pair and has far-reaching implications for both theoretical linguistics and applied natural language processing. By providing a rich, diverse, and meticulously annotated dataset, we enable more robust and naturalistic speech and language technologies for Telugu-English bilinguals.

Moreover, the methodologies developed for this corpus can serve as a template for creating similar resources for other language pairs, contributing to the broader field of multilingual NLP and speech processing. As global communication continues to blur language boundaries, resources like this become increasingly vital for developing technologies that can keep pace with the dynamic nature of multilingual speech.

REFERENCES

[1]. Heike Adel, Ngoc Thang Vu, and Tanja Schultz. Syntactic and semantic features for code-switching factored language models. In IEEE/ACM Transactions on Audio, Speech, and Language Processing, volume 23, pages 431–440. IEEE, 2015.
[2]. Kalika Bali, Jatin Sharma, MonojitChoudhury, and YogarshiVyas. A corpus of english-hindi code-mixed tweets for sentiment analysis. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 2551–2556, 2014. 7
[3]. Amitava Das and Bj¨ornGamb¨ack. Code-mixing in social media text: The last language identification frontier? In TraitementAutomatique des Langues, volume 54, pages 41–64, 2015.
[4]. Braj B Kachru. The Indianization of English: The English language in India. Oxford University Press, 1983.
[5]. Dau-Cheng Lyu, Ren-Yuan Lyu, Yuang-chin Chiang, and Chun-Nan Hsu. Speech recognition on code-switching among the chinese dialects. In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 4350–4353.IEEE, 2010.
[6]. Dau-Cheng Lyu, Tien-Ping Tan, Eng-SiongChng, and Haizhou Li. Language identification in code-switching utterances using syllable structure and phonotactics. Proceedings of Interspeech, pages 3186–3190, 2015.
[7]. Pieter Muysken. Bilingual speech: A typology of code-mixing. Cambridge University Press, 2000.
[8]. Carol Myers-Scotton. Duelling languages: Grammatical structure in codeswitching. Oxford University Press, 1993.
[9]. Kushagra Singh, Indira Sen, and PonnurangamKumaraguru. Language identification and named entity recognition in hinglish code mixed tweets. Proceedings of ACL 2018, Student Research Workshop, pages 52–58, 2018.