Political Discourse Analysis: A Case Study of Code Mixing and Code Switching in Political Speeches

Dama Sravani, Lalitha Kameswari, Radhika Mamidi
Language Technologies Research Centre
International Institute of Information Technology
Hyderabad, Telangana, India
{dama.sravani, v.a.lalitha}@research.iiit.ac.in
radhika.mamidi@iiit.ac.in

Abstract

Political discourse is one of the most interesting data to study power relations in the framework of Critical Discourse Analysis. With the increase in the modes of textual and spoken forms of communication, politicians use language and linguistic mechanisms that contribute significantly in building their relationship with people, especially in a multilingual country like India with many political parties with different ideologies. This paper analyses code-mixing and code-switching in Telugu political speeches to determine the factors responsible for their usage levels in various social settings and communicative contexts. We also compile a detailed set of rules capturing dialectal variations between Standard and Telangana dialects of Telugu.

1 Introduction

Gumperz (1982) defines Code Switching (CS) as the juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or sub-systems. On the other hand, Code Mixing(CM) refers to the embedding of linguistic units such as phrases, words and morphemes of one language into an utterance of another language (Myers-Scotton, 1997). So broadly speaking, CS occurs across sentences/phrases and CM within a sentence/phrases (though some researchers do not distinguish the two).

Gumperz (1982) researched specific speech events to examine the relationship between speakers' linguistic choices. They also looked for CS instances, either between languages or between varieties of the same language, to find out in what situation and with what interlocutors, CS occurs and CS may signal various group memberships and identities. Gumperz (1977) found that local dialect carried great prestige, and as a person's native speech is regarded as an integral part of his family background, a sign of his local identity. However, when interacting with members of other communities and with tourists, the residents would use the standard dialect.

Foster et al. (1981) states that language is not neutral or universal in a political context. Language is used to reflect many historical, cultural and social identities associated with the politician. In a multilingual country like India, CM and CS are a norm. They not only reflect a person's association with more than one language or a dialect, but also conveys their social identity in a given context. In this paper, our aim is to look at CM and CS as two distinctive techniques used for political gain.

The matrix language we have chosen to study these phenomena is Telugu, a South-Central Dravidian language predominantly spoken in India's Southern parts, especially in Andhra Pradesh and Telangana. There are many regional dialects and sub-dialects in Telugu, but the three major dialects are the Coastal Andhra dialect, the Telangana dialect which has a significant influence of Urdu/Dakhni, and the Rayalaseema dialect. The variety spoken by the educated class from the interior districts of Andhra area was modernised and elevated to 'Standard Telugu' status in 1969 and since then has been widely used in textbooks, newspapers and other formal communication. It is also referred to as Modern Standard Telugu (MST) in Krishnamurti et al. (1968).

Even though Telugu is one of India's largest spoken languages with more than 80 million speakers, there is a severe dearth of resources in Telugu, which makes it hard for NLP research. For our purpose, we did not find any corpus for Code Mixing and Switching in Telugu. Hence, we created a corpus of political speeches in Telugu consisting of 1134 sentences and about 10000 words. Since our work is closely associated with a set of rules and statistical observations corresponding to those rules, the corpus size was sufficient to give us good results. We will further examine the factors re-

Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching, pages 1–5
June 11, 2021. ©2021 Association for Computational Linguistics
https://doi.org/10.26615/978-954-452-056-4_001

sponsible for varying levels of CM and CS in these speeches.

2 Related Work

Kameswari and Mamidi (2018) conducted a study about various interpersonal speech choices in election campaign speeches, including the usage patterns of nouns, pronouns, kinship terms, rhetorical questions, etc. There are a few more studies (Martinez Guillem (2009), Ilic and Radulovic (2015), Kampf and Katriel (2016)) which analyse the deeper intention behind the choice of words and phrases using the famous Speech Act theory by Searle et al. (1980) and the Sociocognitive model by Van Dijk (2014).

There has been some work recently on CM and CS involving Indian languages. But most of the work is done in the social media domain and involves Hindi-English pair because of the easier availability of data. Bali et al. (2014) worked on code mixed tweets in English-Hindi. They tried to differentiate between borrowing and code-mixing based on the frequency of co-occurrence of words in tweets.

In Dravidian languages, there is very little work done in this area so far. Srirangam et al. (2019) created a corpus for Named Entity recognition in English-Telugu code mixed tweets,Jitta et al. (2017) created a English-Telugu code mixed conversational data for Dialog Act recognition.

There has been very less work on analysing the Telangana dialect. Bhaskar wrote a book named *Telanagana Padakosam* with Telangana words and their corresponding words in Standard dialect. Also, he has drawn few observations that are common in Telangana dialect.Chakravarthy (2016), An Annotated Translation of Kalarekhalu A Historical Novel by Ampasayya Naveen, describes the important phases that lead to the Telangana state. Few cultural words have been retained without translating into English. Sastry (1987) provided a prosodic analysis of Telangana dialect.

To our knowledge, this work is the first of its kind, which analyses CM and CS together in political discourse along with dialectal level code-mixing analysis. We aim to understand these phenomena as a speech choice and its effect on the audience in politics.

3 Dataset and Annotation

3.1 Dataset collection

Even with the advent of social and print media, in-person modes of communication such as campaigns and political speeches remain the most preferred ways of communicating with the general public for politicians. They try to ensure that the audience feels connected to them, thereby increasing their potential votes. This is done strategically and persuasively.

We chose our speakers as Mr K Chandrasekhar Rao (KCR), the Chief Minister of Telangana and Mr Chandra Babu Naidu(CBN), former Chief Minister of Andhra Pradesh. KCR is the founder of the Telangana Rashtra Samiti (TRS) party and is widely regarded as the face of the Telangana movement for a separate state in 2014. CBN is the leader of the Telugu Desam Party. They use a variety of dialects and languages such as Telangana Telugu, Modern Standard Telugu, Urdu, English and Hindi in their speeches.

We chose a total of 6 speeches of both the speakers in three different social settings and communicative contexts to analyse the levels of code-mixing and code-switching as follows:

1. **Public Meetings in Telangana**: KCR's speech was during the Telangana movement, meant for the creation of a new state. He addressed the pathetic situation of Telangana residents and also discussed the plan and policies for the new state.CBN's speech is during the Telangana elections in 2018. The audience were residents of Telangana. We will refer to this as communicative event 1.

2. **Felicitating Dr. Venkaiah Naidu when honoured as Vice President**: KCR and CBN's speeches were with MLAs and other parliament members of their respective states,*viz.* They spoke about Dr Venkaiah Naidu's great qualities and praised him for his service to the nation and attaining one of its highest positions. We will refer to this as communicative event 2.

3. **Capital Development**: In these speeches, both the speakers were talking about developments of capitals. In the KCR speech, the audience were Government officials and local politicians of Telangana. CBN addressed the

collectors of Andhra Pradesh. We will refer to this one as communicative event 3.

Though the speeches were available on YouTube, none of the existing off-the-shelf speech to text systems could serve to capture the speech effectively along with the dialectal variations in the language. Therefore, we manually transcribed the speeches in the WX format (Diwakar et al., 2010) and verified the transcription with the help of native speakers. The duration of the speeches is 100 minutes for each speaker, and after transcription, it consists of 1134 sentences. The total word count is around 10000.

3.2 Annotation

All speeches are annotated for the usage of CM and CS at the word level. Each speech is annotated for Dialectal level code-mixing(DCM), Language level code-mixing (LCM) and Code-Switching(CS). We will further examine how CM and CS will vary in different social settings and communicative contexts for pragmatic reasons.

3.2.1 Guidelines to handle dialectal level code-mixing

The subjects of our study use Telangana dialect and MST more often compared.

To our knowledge, there has been no exhaustive set of observations differentiating these two varieties Telangana from MST. We took some observations from the book by Bhaskar and Sastry (1987). Few more observations are drawn from texts of Chakravarthy (2016). Also, we compiled a few more observations from a TV news program named *Teenmar news* which uses Telangana dialect. After removing duplicates, we categorised the observations and segregated them into three categories: Vowel rule (V), Consonant rule (C) and the other rules which apply to syllables (S). The rules in each of these three categories are further classified as *Addition, Deletion or Replacement*, based on the kind of operation performed.

We came up with over 50 tags for these observations capturing the pattern differences between the Standard and Telangana dialects. If a word follows any of these observations, then it is marked as 1 under the category DCM. Else, it is marked as 0. In this paper, we present a few observations which are prominent in our data. The writing convention followed is:

[Standard dialect word] - [Telangana dialect word]:

1. **Vowel rules**
   - **Deletion**: In Telangana dialect, vowels are dropped at the end of some words. For example:

     nenu - nen
   - **Replacement**: Long vowels are replaced with short vowels.

     vastAru - vastaru
2. **Consonant rules**
   - **Addition**: In Telangana dialect g is added at the start for few words.

     ippuDu - gippuDu
   - **Deletion**: In some words *v* is dropped at the beginning of the word. This occurs in nouns, pronouns and verbs

     vAna - Ana
   - **Replacement**: Voiced consonants are replaced with voiceless consonants in some words.

     pedda + kAleV- peddagAleV -
     cAlu - jAlu
     peTTAru - beTTAru
3. **Syllable rules**
   - **Deletion**: Dropping of the syllable which precedes the /d/ sound. In some cases, after the dropping, the preceding vowel is lengthened. This is mostly observed in terms associated with spatial deixis.

     ikkaDa - IDa
   - **Replacement**: For the verbs in past tense, The second last syllable's long vowel gets replaced with *in/i/shortening of vowel/ina*. These are further subcategorised based on gender, number and person.

     cesAru - jeSinru
     cesAvA - jeSinavA
     cesAru - jeSiru

3.2.2 Guidelines to handle language level code-mixing

In our paper, language level code-mixing is said to occur when two or more languages or language varieties are used at a morphological level. To be more precise, it occurred when English root words were suffixed with Telugu plural markers, and morphological suffixes in one word or English/Hindi words are used.

pArtIlu - party + lu
kAlejIlo - College + lo
rejiyanga - region + ga

If a word follows these observations, then it is marked as 1. Else, it is marked as 0 for language level code-mixing.

3.2.3 Guidelines to handle code-switching

All the language variations at the sentence level, i.e. if the sentence or phrase with more than one word is in a different language, then it is considered under code-switching. Here as our speeches are in Telugu, sentences or phrases in languages other than Telugu come under this category. All the words in these sentences/phrases are marked as 1.

mIru ganaka commitement won
tIskunte, Yes sir come on let us move annAru

In the above sentence, all the words in the phrase *Yes sir come on let us move* are marked as 1 under the category code-switching.

4 Observations and Results

After annotating based on these guidelines, the results are tabulated as follows.

Speech	No.of Words	Dialectal level Code-mixing	Language level Code-mixing	Code-Switching
1	2153	19.9%	E-5.4% H-0.8%	E-0.1% H-17.4%
2	1137	12%	E-3.1% H-0%	E- 2.1% H-0%
3	2654	15%	E-9.2% H-0%	E-9.9% H-0%

Table 1: KCR Speech Statistics (E-English, H-Hindi/Urdu)

Speech	No.of Words	Dialectal level Code-mixing	Language level Code-mixing	Code-Switching
1	1357	8.91%	E-4.64% H-0%	E-1.76% H-0%
2	1960	3.82%	E-4.7% H-0%	E- 4.33% H-0%
3	984	2.7%	E-7.01% H-0%	E-39.63% H-0%

Table 2: CBN Speech Statistics

In communicative event 1, as they were addressing Telangana residents, relatively higher levels of Telangana dialect are observed in speeches by both the speakers to get more *connection with the audience*. However, KCR has used more Telangana dialect in his speech than CBN. KCR was fighting for a separate Telangana state. CBN speech was during the Telangana elections in 2018. His ideology doesn't align with KCR. In addition to connection with the audience, *ideologies of the speaker* also impact the levels of code-mixing and code-switching. KCR also uses high levels of code-switching in Hindi for establishing a better connection with the audience as the Telangana dialect is influenced by Hindi/Urdu.

In communicative event 2, KCN and CBN addressed MLAs and other parliament members of Telangana and Andhra Pradesh. In CBN's speech, the usage of MST can be due to the absence of Telangana residents. However, in KCR speech, most of them are Telangana residents, yet lesser levels of Telangana dialect are observed. So, *context of the speech* also determines the levels of code-mixing and code-switching. In this communicative event, as they were addressing a national topic, MST, lesser language level code-mixing and lower code-switching levels are observed.

In communicative event 3, English usage is high in both speeches than other speeches as the meeting is about capitals and all government officials may not be aware of the local language. In KCR speech, local politicians are also part of the meeting, so Telangana dialect usage is prominent. Whereas in CBN speech, very high levels of English is used as the meeting is only with collectors.

5 Conclusions and Future Work

In this paper, we looked at the phenomenon of CM/CS between dialects of Telugu, MST and languages like English and Hindi/Urdu for different communicative contexts. The audience, ideologies of the speaker and context of the speech impacted the speakers linguistic choices.

Our transcribed and annotated speeches[1] can be further be used to develop dialectal speech recognition systems.We present a very detailed set of observations and annotation guidelines to capture the dialectal variations between the MST and Telangana dialect of Telugu. These could be studied and extended to handle dialectal variations in other languages, especially Dravidian languages like Tamil and Kannada. These can also help develop Machine Translation systems equipped for several dialects within a given language pair. Further, we would like to expand our data and examine other factors responsible for code-switching and code-mixing.

[1] https://github.com/damasravani19/CodeMIxingCodeSwitchingInPoliticalSpeeches

6 Acknowledgements

We thank SriHarshitha Bondugula for aiding us in transcribing political speeches in the Telangana dialect.

References

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. “i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 116–126.

Nalimela Bhaskar. *Bhaskar,*.

I Pavan Chakravarthy. 2016. *An Annotated Translation of Kalarekhalu A Historical Novel by Ampasayya Naveen*. Ph.D. thesis, The English and Foreign Languages University, Hyderabad.

Sapan Diwakar, Pulkit Goyal, and Rohit Gupta. 2010. Transliteration among indian languages using wx notation. In *Proceedings of the Conference on Natural Language Processing 2010*, CONF, pages 147–150. Saarland University Press.

Leslie D Foster, Dennis J Gallant, William D Drew, and Cecil R Lohrey. 1981. Columnar patient care service facility. US Patent App. 06/004,211.

John J. Gumperz. 1977. The sociolinguistic significance of conversational code-switching. *RELC Journal*, 8(2):1–34.

John J Gumperz. 1982. *Discourse strategies*, volume 1. Cambridge University Press.

Biljana Misic Ilic and Milica Radulovic. 2015. Commissive and expressive illocutionary acts in political discourse. *Lodz Papers in Pragmatics*, 11(1):19.

D. S. Jitta, K. R. Chandu, H. Pamidipalli, and R. Mamidi. 2017. “nee intention enti?” towards dialog act recognition in code-mixed conversations. In *2017 International Conference on Asian Language Processing (IALP)*, pages 243–246.

Lalitha Kameswari and Radhika Mamidi. 2018. Political discourse analysis: A case study of 2014 andhra pradesh state assembly election of interpersonal speech choices. In *PACLIC*.

Zohar Kampf and Tamar Katriel. 2016. Political condemnations: Public speech acts and the moralization of discourse. *The handbook of communication in cross-cultural perspective*, 312:324.

B. Krishnamurti, P.S. Sarma, and K. Civam. 1968. *A Basic Course in Modern Telugu*. sole distributors Motilal Banarsidass, Delhi.

Susana Martinez Guillem. 2009. Argumentation, metadiscourse and social cognition: organizing knowledge in political communication. *Discourse & Society*, 20(6):727–746.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

J. Sastry. 1987. A study of telugu regional and social dialects : a prosodic analysis.

John R Searle, Ferenc Kiefer, Manfred Bierwisch, et al. 1980. *Speech act theory and pragmatics*, volume 10. Springer.

Vamshi Krishna Srirangam, Appidi Abhinav Reddy, Vinay Singh, and Manish Shrivastava. 2019. Corpus creation and analysis for named entity recognition in telugu-english code-mixed social media data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 183–189.

Teun A Van Dijk. 2014. *Discourse and knowledge: A sociocognitive approach*. Cambridge University Press.