Mid-Project Report
LTU

Tanishq Chaudhary, 2019114007
Mayank Goel, 2019114004
Shivansh Subramaniam, 2019114003

Introduction

Code Mixing is a linguistic phenomenon where multilingual speakers mix characteristics and/or vocabulary of different languages together.

In our Project, we have worked with English-Hindi Code-mixed data, taken from natural sources (and thus, in the Latin script). At this current stage, our end-goal isn't clear, however the motivation is to see how linguistic analysis in code-mixed data will work, if we approach it as a single language.

Code-Mixed Constructions

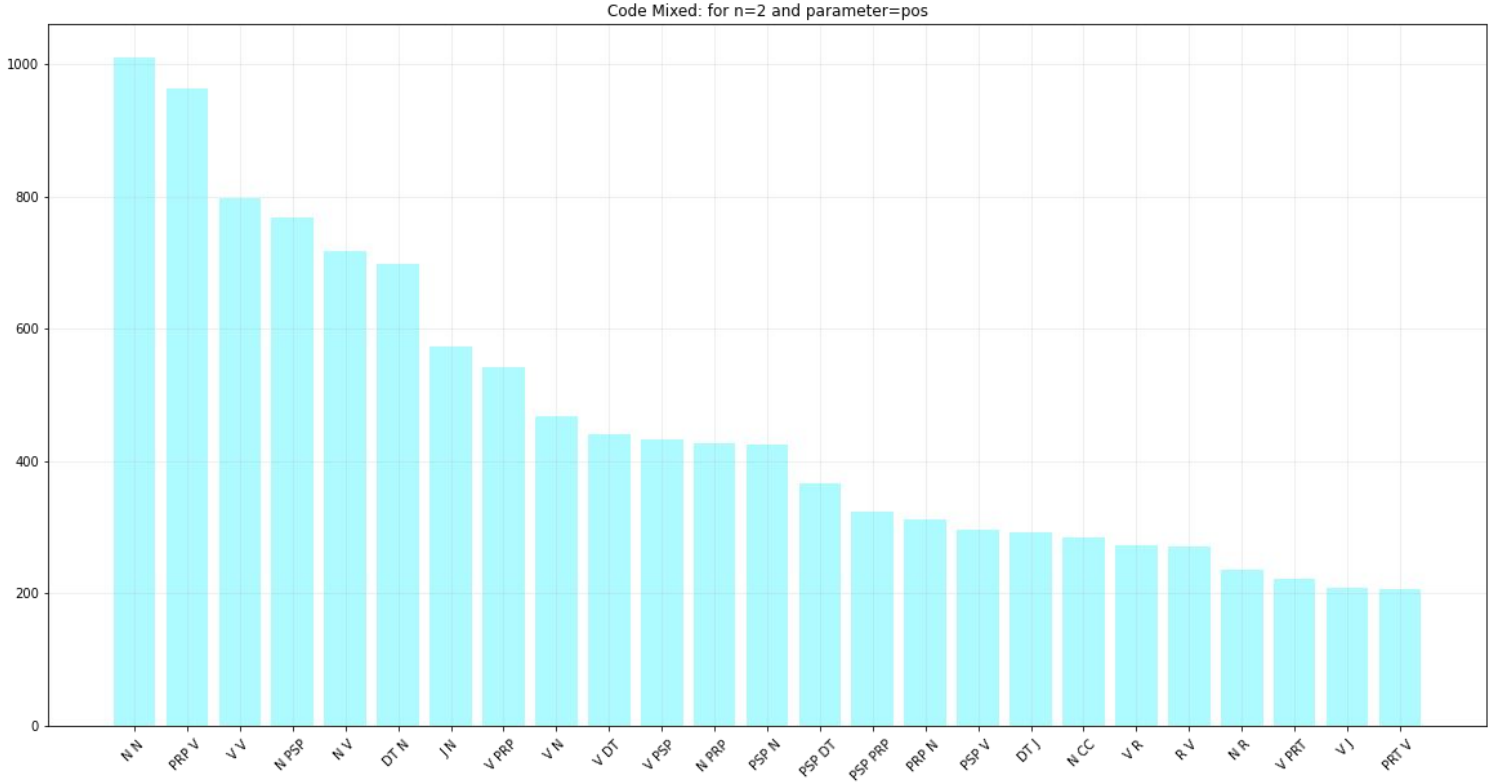
1. Yogarshi Vyas, S Gella, J Sharma, K Bali, and M Choudhury. 2014. POS Tagging of English-Hindi Code-Mixed Social Media Content. In Proc. EMNLP
2. Code-mixing, language variation, and linguistic theory:: Evidence from Bantu languages by Eyamba G.Bokamba
3. Uncovering Code-Mixed Challenges: A Framework for Linguistically Driven Question Generation and Neural based Question Answering by Deepak Gupta, Pabitra Lenka, Asif Ekbal, Pushpak Bhattacharyya

Dataset

Source :
https://github.com/Mysterious19/POS-tagging-Codemix

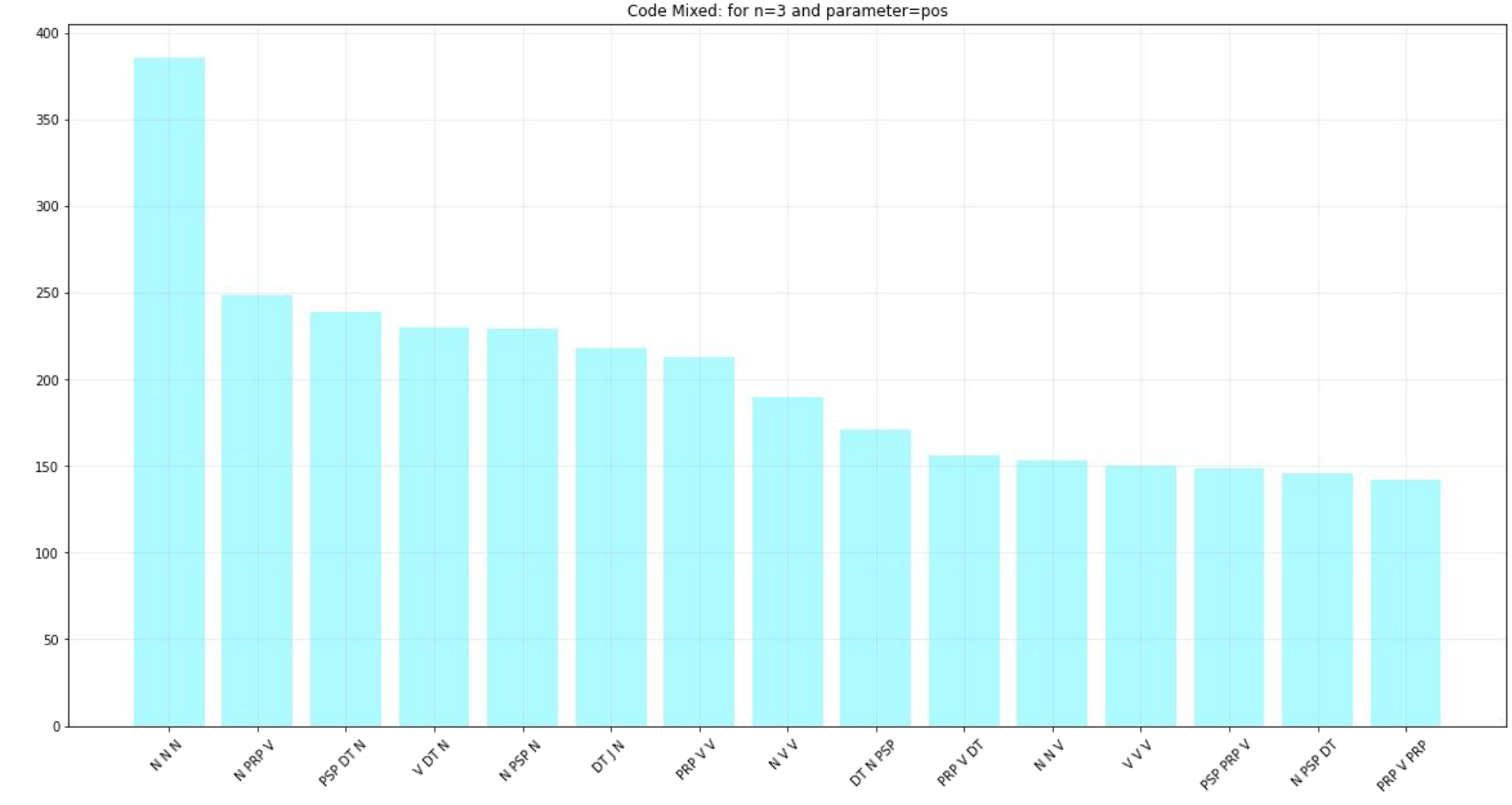
The text is from Whatsapp/Facebook/Twitter in the Latin script, in Hindi-English codemixed language.

This data has been manually tokenized and POS Tagged The quality of POS Tagging is fairly poor, but is sufficient for our analysis.


Code Mixed: for n=2 and parameter=pos
1000
800
600
400
200
0
N N
PRP V
V V
N PSP
N V
DT N
J N
V PRP
V N
V DT
V PSP
N PRP
PSP N
PSP DT
PSP PRP
PRP N
PSP V
DT J
N CC
V R
R V
N R
V PRT
V J
PRT V

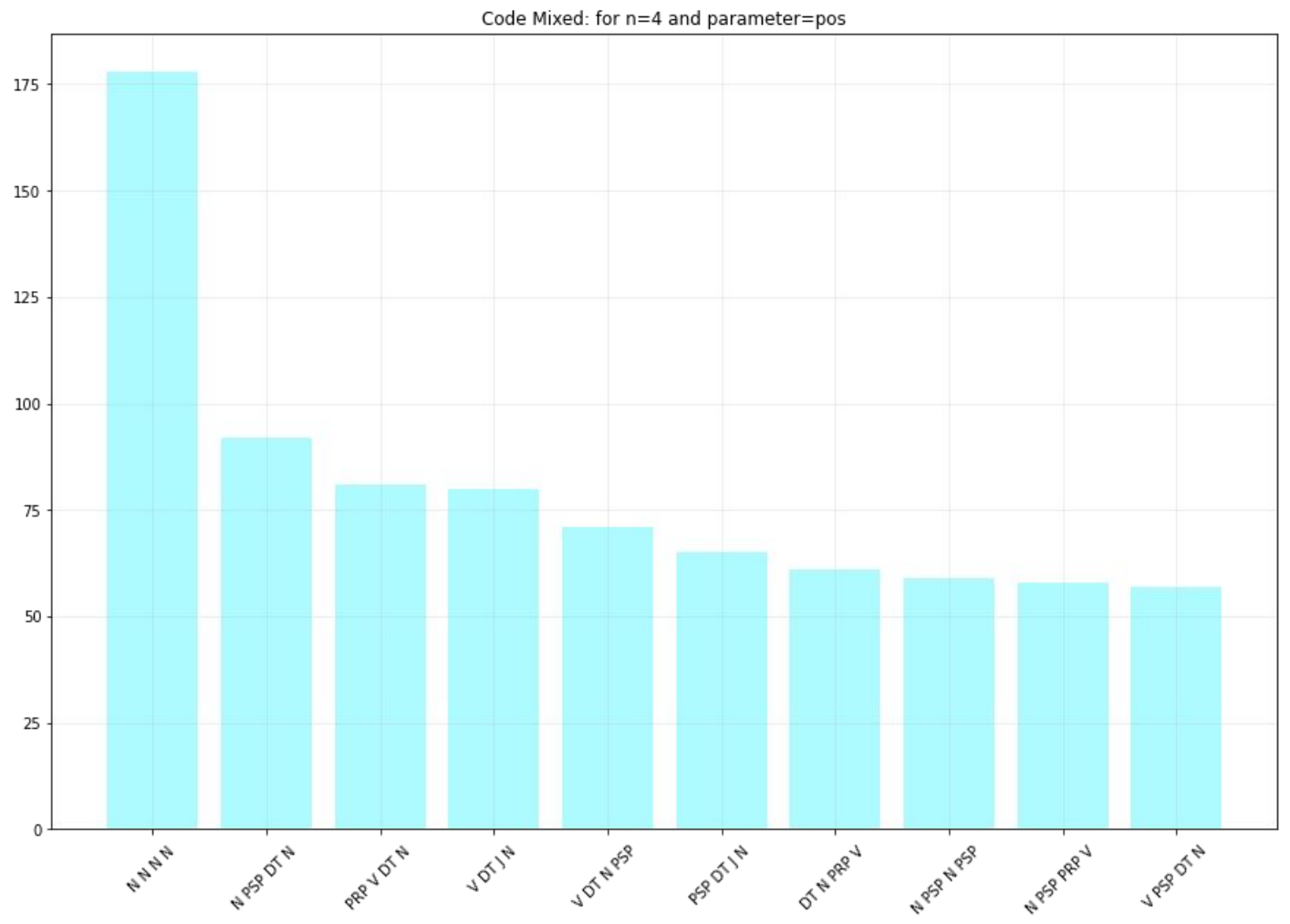
POS: n = 2

- Here, we start to see patterns in our data, which would be very helpful for future analyses.
- We see for example the occurrence of N N in large amounts
- V V is also present, which is perhaps best explained by Hindi's verbs, which are followed by multiple auxiliaries. (The current labelling uses it that way)
- We also see J N, which hints at both the languages following the structure of Adjective-Noun


Code Mixed: for n=3 and parameter=pos
400
350
300
250
200
150
100
50
0
N N N
N PRP V
PSP DT N
V DT N
N PSP N
DT J N
PRP V V
N V V
DT N PSP
PRP V DT
N N V
V V V
PSP PRP V
N PSP DT
PRP V PRP

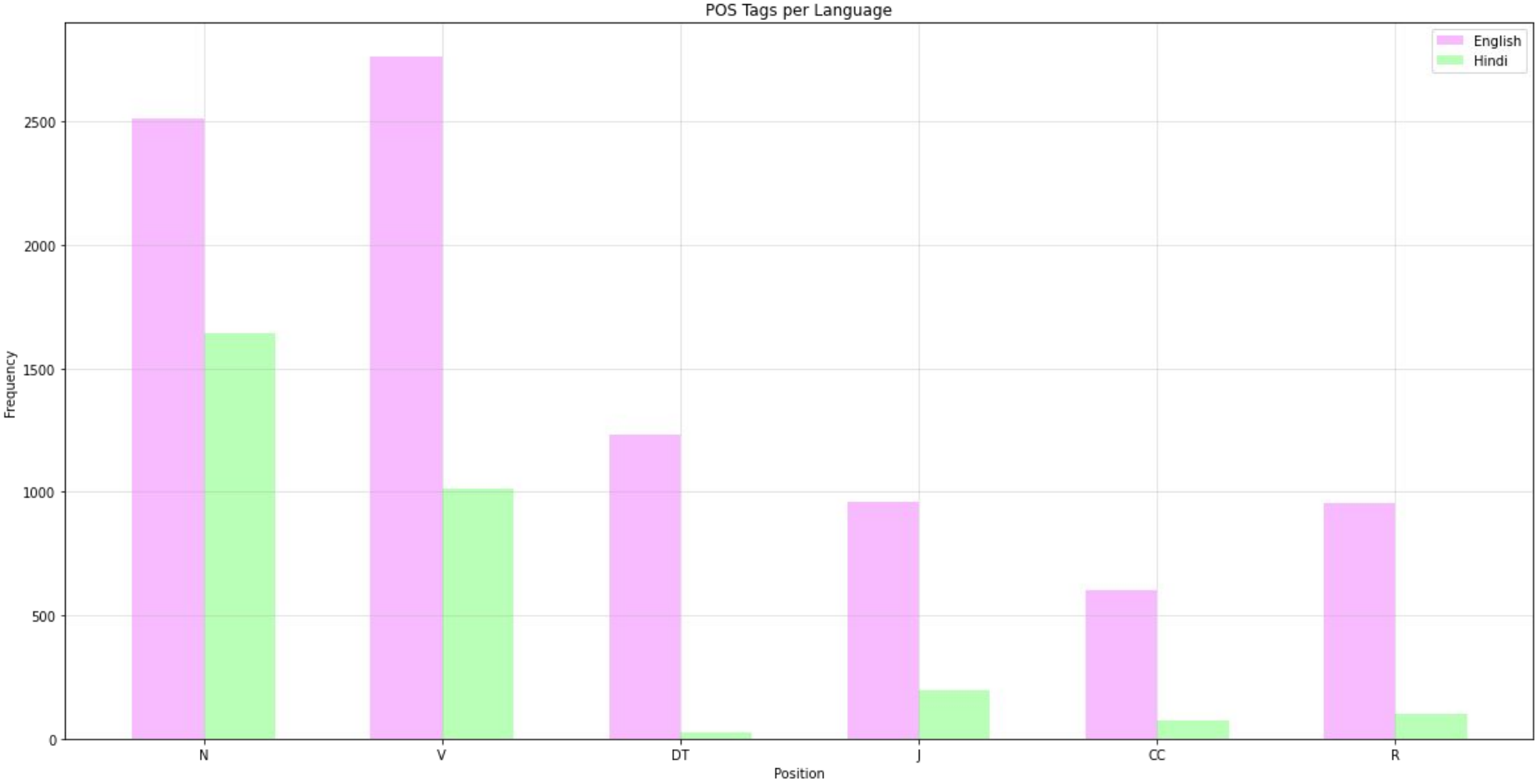
POS: n = 3

- We see the use of V DT N, which hints at the larger presence of English, having the order SVO, instead of Hindi's SOV.
- We also have instances of N N V, affirming that even code-mixed constructions do not break the order of either the language; since both are again, SVO or SOV
- We also see V V V, which is as we predict, due to Hindi's auxiliaries.


Code Mixed: for n=4 and parameter=pos
175
150
125
100
75
50
25
0
N N N N
N PSP DT N
PRP V DT N
V DT J N
V DT N PSP
PSP DT J N
DT N PRP V
N PSP N PSP
N PSP PRP V
V PSP DT N

POS: n = 4

- We see N N N N, which is not expected in normal use of the langage. This is one of the places where the tagger used by the dataset author, breaks down.
- We see N PSP N PSP, which points to the Hindi's use of karakas to assign roles to the nouns.
- We see V DT J N. Already assuming a larger influence of English, we can see the noun-phrases coming to light, with the structure only found in English - due to the presence of DT


POS Tags per Language
English
Hindi
2500
2000
1500
1000
500
0
Frequency
N
V
DT
J
CC
R
Position

POS: English vs Hindi

- One big difference we can immediately see is any lack of DT in Hindi. English uses a lot of articles in comparison.
- However, if one looks closely, we can see that English is proportionately higher than Hindi in all the major aspects - be it nouns or verbs. This re-affirms our initial guess of most sentences of the code-mixed data embedded in English, as the language that provides the majority structure to such constructs.

Future Work

We propose further work expanding on the exploration of multiple languages in India. Here, we will gather data (possibly using twitter) of various codemixed tweets. We can then compare which english words/phrases have a higher incidence common between all languages, which specific phrases/words are higher in each language (and a qualitative analysis of both), degree of code mixing and their variance between languages.

Challenges Faced

There are some core challenges we have faced:

- Lack of POS Tagged data for codemixed corpus
- The data we got had numerous inconsistencies, which labelled PRP as pronoun and PRP as particle for hindi.
- (Lots of) Inaccurate or incomplete tools for language identification.
- Dearth of existing rule-dependent research on code-mixed constructions.

Challenges Faced

- (Prospective) Limited Proficiency by Team Members for Bengali/Tamil, not suitable for extensive annotation.
- For any deeper analysis, we would need shallow parsing, however, tools we found all had low accuracies; not enough for drawing valid inferences.

Thank You!