# Feminism, Jaat, aur Code Mixing

Niyati Bafna and Preetha Datta

## I. Introduction

The objective of this project is to analyze the relative presence of code-mixing in the lingo that has developed on social media, specifically, Twitter, around two major sociocultural discourses in India: that on caste-concerned issues, and that on gender concerns[1].

The presence of code-mixing in a particular instance of speech/text in a bilingual society is contingent on several factors: for example, the proficiency of the particular speaker, the social situation s/he is in, a tendency to accommodate the conversant, etc. (Bali 1617). Suppose that we have accounted for variations in these, and observed the following two utterances:

1) Vah ladki chair par khadi ho gai
2) Yeh mera personal maamla hai.

Both instantiate code-mixing on a lexical level. There is a difference, however, in the fact that one might reasonably hear

1) Vah kursi par khadi ho gai
   whereas, it is much more anomalous for a modern Hindi speaker to say
2) Yeh mera vyaktigat maamla hai

Putting it differently: were we to ask 1000 speakers independently to express each of these two thoughts, we might expect to find that the percentage of people who preferred 'personal' to 'vyaktigat' would be higher than that of those who preferred 'chair' to 'kursi'. Why might this be?

One explanation is that the discourse of privacy and the personal has been culturally borrowed from a Western ideology. It is not, of course, an imported technology like a 'phone' or a 'fridge'; Hindi might well contain the vocabulary to support it ('niji' maamla, 'apna' maamla), but using the original English code 'personal' might remain the more popular way of expressing a concept that was introduced from outside the cultural currency of Hindi. This is to say that the topical features, or the local semantic properties, of a discussion influence the lingo in which it occurs, owing to the socio-cultural history of the discourse itself.

## II. Hypothesis

**The relative presence of Hindi in the topical category of gender issues, including but not limited to activism and protest, will be less than that in the topical category of caste-based issues, including but not limited to reservations, Dalit campaigns, and activism**[2].

The idea of feminism as we comprehend it today emerged in India post- independence. It is still regarded as an upper-class, upper-caste movement, ideologically populated, perhaps, largely by the English-educated stratum of India. On the other hand, there is reason to think that an engagement with caste politics is more deeply entrenched in Indian society and discussion, since it often forms a nerve center for nationwide political discussion, often conducted in regional languages. In both movements, speaking English itself is a political move, since English is the language of the 'liberated' and upward mobile in India.

## III. Methodology

This project, drawing upon the annotation scheme as suggested by Bali et al., will fix the modality, i.e. the domain, to be Twitter, the discursive dimension, i.e. the nature of the speaker-audience, to be multiparty or public, and the social hierarchical dimension to be informal, as is the nature of the Twitter platform. We will then identify tweets that belong to either of the discourses under discussion in an annotated corpus, with the help of a predetermined list of associated words and their translations[3].

We performed the following analyses :

1) Analysis of the occurrence of a list of ideological words, appearing in their Hindi and English equivalents, to determine the trends in ideological discussion.
2) Word-based analysis of English-Hindi proportions
3) Sentence-based analysis of English-Hindi proportions.
4) Measure the run length of CS when it exists.
5) Identify tweets with CS.

**Therefore, our methodologies, indexed according the above list is as follows:**

0) The number of tweets we get for each is a general indicator of the relative strength and popularity of these two conversations on Twitter in India
1) The list of ideological words (provided below) contains two sets: one of Hindi equivalents and one of English equivalents. We measured the occurrences of each, and compiled the results which will be shown in the following sections.

[1] These terms are vaguely put on purpose; they will be more rigourously defined.

[2] This is assuming that English is the unmarked language on Twitter. If not, that does not affect the nature of the analysis.

[3] The segregation of these discourses is not to ignore the possibility of intersectionality between these movements; these will be handled as instances of both.

2) This can be performed on the raw corpora.

For the next three analyses, we had the corpora tagged by the Microsoft LID. We also made certain assumptions:

*1)* ***Assumption:*** There are very few English sentences written entirely in the *Devanagri* script, (although we might have long segments of Roman transliterated Hindi). This seems true from observation in the corpora – it is also fairly intuitive, since the English – Hindi Roman – *Devanagri* baseline for Twitter is 70-15-15; it is probably rare that people typing in the *Devanagri* script are expressing themselves in English. In any case, we can observe our own corpus to validate this assumption. We assumed that any sentence in the Devanagari script, therefore, is a Hindi sentence.

*2)* ***Assumption:*** There is very little script-mixing in one sentence. Again, this is true from observation. People do insert English acronyms (SC, BJP, etc.) into their *Devanagri* tweets, but they do not switch scripts halfway through a sentence, or for particular phrases. Note that we are not assuming that tweets cannot contain different scripts: this is both possible and observable. It is only at the sentence level that this assumption holds.

The necessity for the above two assumptions arise from the limitation that we do not have an en-hi LID software for *Devanagri* script; we cannot analyse the language of the words in a Devanagari script sentence. However, there is reason to believe that these assumptions will only result in marginal error.

In general, we assume that sentences are either in Roman or in *Devanagri* script, with perhaps some lexical level insertions in the other script, which we can ignore. Furthermore, a sentence in *Devanagri* script is a Hindi sentence, although it may contain some transliterated English words.

Now, we LID the corpus, in which: a) the beginning of each new tweet is marked b) The beginning of each new sentence is marked.

We counted the number of Devanagari sentences in the corpus. These are Hindi tweet-sentences. They contain some CS, of course, but we do not at present have the resources or the scope to measure this.

*3)* ***Assumption:*** The language of a sentence is usually that in which the larger number of words is present. This tends to be true in general, but not, of course, always – e.g. in the sentence “She told me to come but I said that Hindi phrase HP” the sentence is English, no matter how long HP is. However, in most cases, this assumption holds.

Given the LID‘d corpus, we do a hi-en comparison for each Roman script sentence, and assume its language is that which has greater word presence. This gives us a sentence-based analysis, i.e. analyzing at the lowest unit of a conversation, and ignoring CS for the moment. The reason we are choosing the sentence as a unit and not a tweet is because a multi-line tweet may contain two lines, say, in different languages, in which case the dominant language cannot be considered at that containing the most words (in fact, we may say that there is no such language). That is, we are trying to analyse, simplistically, the language in which this conversation is occurring at sentence level.

3) Sentence-based analysis: This analysis can only be performed on the LID corpus, since we cannot identify English *Devanagri.*
4) Code-switched tweets: Once again, we cannot determine this heuristic exactly, for the same reason. If a (multi-line) tweet contains English and Devanagari script, we may safely assume it contains code-mixing. We assume that purely Devanagari tweets are all-Hindi.
5) CS run-lengths: Given a line, we can find the run length of CS, if present. We have labelled the language of the line in (2). Typically, a run-length of 3 or more in the opposite language indicates phrasal CS. This figure gives us greater indication than (4) as to how fluid the parlance is between Hindi and English i.e. how comfortable are people switching back and forth mid-sentence in this discourse, as compared to between sentences. Of course, this count is a subset of (4). Similarly, we can count CS fragments of different lengths: a 1-length fragment, for example, might indicate lexical mixing. We will discuss this more fully below.

## IV. Constructing the Corpora

We have run 22,307 two-term combinations of words, generated by 271 keywords associated with a conversation around Dalit issues **Dalit corpus** . For the feminism corpus **Feminism corpus**, we had 293 keywords, and generated 23,844 combinations. These keywords were divided in equal proportions between Hindi and English for either corpus – the entire lists are available in **WordSearchResults** . Each combination collected tweets from Twitter with a cap of 30,000 items, which we may safely assume suspect was far over the number of relevant tweets required [4].

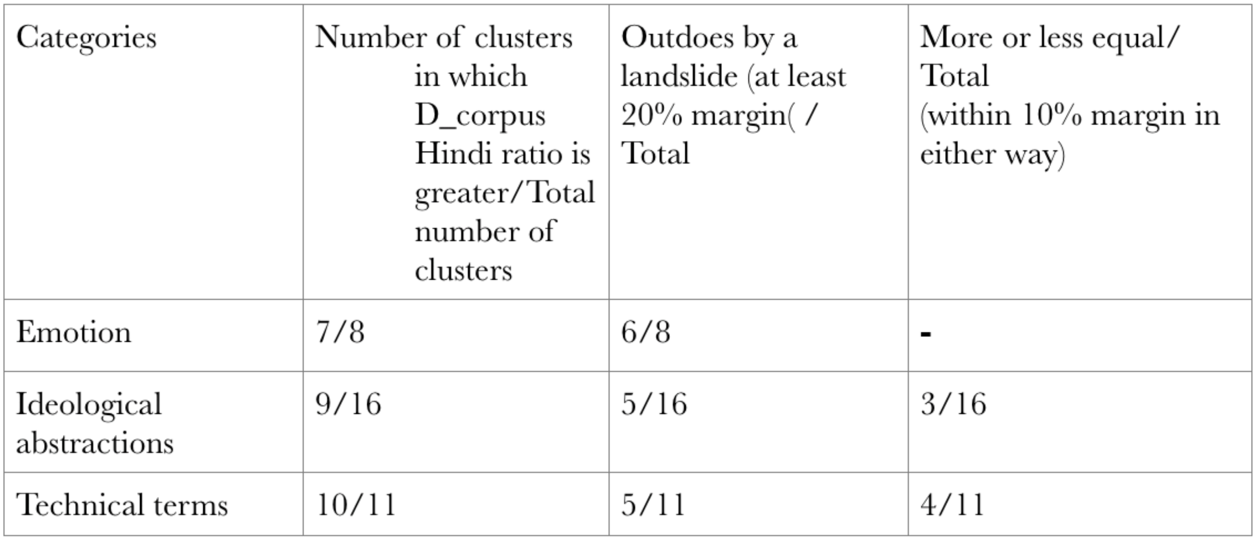
We have tried, in short, to minimize any possible bias on our part, either language-wise or corpus-wise, in the collection of data by making both sets of keywords equally strong with respect to both Hindi and English, and Dalit and Feminism [5].

## V. Word Search Analysis

We have created a list of Hindi-English equivalent sets of key terms that form the centres of discussion with regard to the politics of inequality. These have been segregated

[4] We may assume this because there was no significant difference in time or number of tweets collected per search combination upon experimentally reducing this number, down even to 10,000.

[5] The sheer size of the final combination set makes it unlikely that some negligence or ignorance on our part with respect to one or two words will severely skew the results. In any case, we have tried to be as thorough as possible.



| Categories | Number of clusters in which D_corpus Hindi ratio is greater/Total number of clusters | Outdoes by a landslide (at least 20% margin( / Total | More or less equal/ Total (within 10% margin in either way) |
|---|---|---|---|
| Emotion | 7/8 | 6/8 | - |
| Ideological abstractions | 9/16 | 5/16 | 3/16 |
| Technical terms | 10/11 | 5/11 | 4/11 |

Figure 1: TABLE 1.

by a rough semantic/functional categorization: **Emotion-denoting**, **Ideological abstractions** and **Technical terms**.

This list is amenable to expansion and refinement, of course. At the moment, we have included only the broadest of terms, and its size is as following:

1) Emotion : 8
2) Ideological abstractions : 16
3) Technical terms : 11

Each of the words in the above list is searched in a cluster, i.e. along with its close synonyms, possible spelling variants, etc. so as to gather as accurate an estimate of how many times the concept appears in the corpus as possible.

These terms are not specific to either corpus or discussion (except ‘feminism’ and ‘untouchability‘, belong to the last list). In any case, we are looking for a ratio of Hindi: English from each corpus, rather than the absolute count of occurrences. This tells us about the language that people prefer to use within these arenas, and which language the ideological, technical, and emotional thrust of the conversation lies, no matter the ‘base‘ language or the surrounding words.

**Search List**[6] **:**

1) Emotion
    - Anger,गुस्सा , raag, gussa
    - Happy, Happiness, खुश, खुशी, khushi, khush
    - Humiliation, शर्म, नीचा , sharm, neecha
    - Sad, sadness,दुखी, दुख , dukhi, dukh
    - Love, affection, romance,प्यार, प्रेम, मुहब्बत, इश्क, pyaar, prem, muhabbat, ishk
    - Empathy, sympathy, solidarity,दया, हमदर्दी, daya, hamdardi
    - Pride, गर्व, मर्यादा , garv, maryada, maryaada
    - Endurance, सहनशीलता , sahanshilta
2) Ideological abstractions
    - Humanity, इंसानियत , insaaniyat
    - Equality,बराबरी, समानता, baraabari, samaanta
    - Inequality, असामता, भिन्नता , asamaanta, bhinnta
    - Freedom, liberty,स्वतंत्रता , swatantrata
    - Upliftment, alleviation,उत्थान , utthaan, uthan
    - Justice,न्याय , nyay
    - Injustice,अन्याय, बेइंसाफी
    - Social, society,समाजिक, समाज, samajik, samaj, samaaj
    - Oppression, repression,दबाव , dabav
    - Abuse, atrocity, atrocities,अत्यचार, हमला ,
    - Discrimination, repression,भेदभाव, पक्षपात, bhedbhav, bhedbhaav, pakshpat, pakshpaat
    - Empowerment, सशक्तिकरण , sashaktikaran
    - Privilege,विशेषाधिकार , visheshaadhikar, visheshaadhikaar
    - Violence, violent,हिंसा, हिंसक
    - Strength,शक्ति, ताकत
    - Progress, advancement,विकास, vikas
3) Technical terms
    - Democracy, लोकशाही
    - Election,चुनाव
    - Government,सरकार
    - Activism,सक्रियता , sakriyata
    - Representative, प्रतिनिधि , pratinidhi
    - Revolution, protest, campaign,आंदोलन, विरोध, संघर्ष, मोर्चा , morcha
    - Religion,धर्म, dharma
    - Rights, अधिकार, adhikar, adhikaar
    - Casteism, untouchability,छुआछूत, मनुवाद, जातीवाद, ब्राह्मनवाद, chuachut, manuvad, jaativad, brahmanvad

[6] Some transliterations/variations, especially Roman transliterations of Hindi words, have not been included over here; they are present in the comprehensive search list, available in WordSearch

- Feminism, नारीवाद , naarivad, narivad
- Law, constitution, कानून, संविधान

### Results for Word Search Analysis

We observe that the percentage of Hindi used in the Dalit corpus to represent these word clusters outdoes that of the Feminism corpus in the majority of words in each category. (Results in Table 1.)

A more detailed representation of the results, with the figures for each word cluster and its constituents and percentages is available in WordSearchResults. This observation aligns with our hypothesis: indeed, **it is commoner for people in the Dalit discourse to express key concepts in Hindi than for people in the Feminism corpus**, no matter what the surrounding language of the tweet is.

### Some Sub-results

The absolute values of the number of occurrences of each constituent in the above clusters are also telling, as well as the individuals percentages.

### Absolute distribution of negative emotions in Feminist corpus

The Hindi/Total ratio for Emotion keywords were consistently higher than 50 percent for Dalit corpus, whereas they varied for Feminism corpus. Interestingly, Feminism corpus hits 98.8 percent for (3), i.e. humiliation-शर्म. While happiness, sadness, love and pride appear about 80percent of the times in English, humiliation and anger are almost always expressed in Hindi in Feminism corpus.

The case of ‘pride‘ is a little different: while both the Hindi figures are high (97.2 percent and 65.5percent for Dalit corpus and Feminism corpus respectively), this is for different reasons. In Dalit corpus, the Hindi sense गर्व (a sort of positive-pride) populates the figure, with 1407 hits, whereas in Feminism corpus, मर्यादा populates it. This signifies, actually, male-pride, from an Indian patriarchal tradition; it not only does not possess an exact English equivalent, possibly promoting Hindi usage, but also contains highly negative connotations in this conversation, fitting in, therefore, with the above Hindi trope of humiliation and anger in Feminism corpus.

### Ideological Abstractions

Here, we are observing ideological words from an essentially Western discourse of humanity, discrimination, equality, and empowerment. (This is not to say that these concepts do not exist in India or Hindi, only that, since the global conversation around them solidified with the likes of the UNHRC, they are naturally embedded in a Westernized discourse stylistic.) We see that in Feminism corpus, where we had expected the conversation to align more with the global discourse, the Hindi percentages are accordingly consistently low, under 10 percent for 7/16 words, and hit 0 percent for ‘inequality‘. The figure spikes for शक्ति, probably because of the slogan of नारी शक्ति (literally, women strength) in Hindi. The word ‘privilege‘ is a prime example of an intensively Western discourse-point: the discussion of privilege politics is ubiquitous in any ‘educated‘ discussion these days. The figures, again, corroborate (only 0.41 percent, 3.33 percent Hindi respectively).

Some other caveats are that certain words, like ‘society’ and ‘discrimination’, show markedly different values for Dalit corpus and Feminism corpus (83.8 percent, 24.6 percent and 56.6 percent, 5.77 percent); this is to say that this ideological conversation of community and rights is present in Hindi, but that while the Dalit conversation chooses to occupy the space of ′भेदभाव ’ and ′समाज ’, the feminism conversation is still inspired mainly by the global human rights registers.

There are some exceptions to the trend: surprisingly, cluster (16) showed near-complete Hindi percentages: 95.3 percent, 82.9 percent, respectively. The Dalit corpus figure may be explained by the politicization and Hindi-ization of this discourse (as we discuss soon), but it is difficult to pinpoint a reason behind the Feminism corpus figure.

### And finally, Technical Terms

The Hindi percentages here are high-ish, both for Dalit corpus and Feminism corpus, which is surpising given that the language of legality and jargon is shifting to English. However, there is an observable trend: we note that more functional words such as ‘government-सरकार′, ‘representative−प्रतिनिधि ’ and ‘election-चुनाव′ are dominated significantly by Hindi in both corpora, while the figure drops to below 20 percent for ‘democracy-लोकशाही′ and ‘activism−सक्रियता ’, which are more conceptual and less likely to appear in, say, a mass discussion about the coming elections. We achieve a rough 50-50 for both corpora for ‘religion-धर्म′ , which is suggestive that religion is still based in the native tongue, even while the surrounding ideological discourse may be tilting towards English.[7] (Of course, while the above observations have been generalized roughly for both corpora, the Hindi figures are still consistently higher for Dalit corpus, as we primarily wished to demonstrate.)

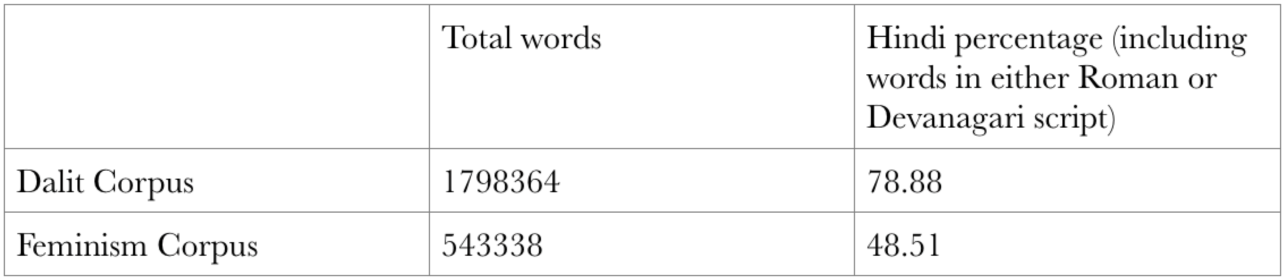
### Intersectionality

The terms ‘feminism-नारीवाद′ and ‘casteism- जातीवाद′ were included in the list as a bit of an afterthought: a look at these figures gives us a snapshot indication of the state of intersectionality in each discourse. Unfortunately, it is clear that intersectionality is near-absent: the feminism cluster appears 11 times in Dalit corpus as compared to 3083 appearances of the casteism cluster, and the casteism cluster appears 24 times in Feminism corpus as compared to 470 appearances of the feminism cluster.

## VI. Word-count analysis

This give us an idea of the landscape before we move into code-mixing: i.e. what are the percentages of Roman English, Roman Hindi, Devanagari Hindi words we are

[7] We are saying this in anticipation of the results in the following analyses, that demonstrate the high amounts of English sentences in Feminism corpus



| | Total words | Hindi percentage (including words in either Roman or Devanagari script) |
|---|---|---|
| Dalit Corpus | 1798364 | 78.88 |
| Feminism Corpus | 543338 | 48.51 |

Figure 2: TABLE 2.

looking at in each corpus. It also helps us understand the significance of our results in (1).

This preliminary heuristic sets the platform for the results that follow. Furthermore, we now have a baseline to juxtapose our results of the WordSearch Analysis against i.e. if roughly 80 percent of words in the corpus are Hindi, then it is all the more important that the word cluster for ‘justice’ is expressed only 24.88 percent times in Hindi. Similarly, given that only 48.51 percent of words are in Hindi in Feminism Corpus, we must be very surprised at the 98 percent figure we get for ‘humiliation-शर्म ’. We can put, thus, each figure that we arrive at in our Word Search Analysis into perspective.

The ratio of the total words in the corpora (about 3:1 in favour of Dalit Corpus) also puts the absolute values of the occurrences of our key terms in Word Search Analysis into perspective. For example, the word clusters for ‘justice’, ‘empowerment’ and ‘rights’ appear in absolute terms more in Feminism Corpus than in Dalit Corpus, despite the 3:1 ratio in the opposite bias. This ties in, once more, with our hypothesis above, about the globalized Englishized discourse we can observe in Feminism Corpus.

Results recorded in Table 2.

## VII. Sentence-based analysis

Now we move on to the sentence; this is different from the analysis of a tweet, because tweets can be and often are multiline, where the user switches languages or scripts for, say, emphasis, or convenience, or any other reason.

What we are trying to analyze via this heuristic is the nature of the conversation of these discourses; for the purpose of this sub-analysis, we treat each sentence as a unit of this conversation, regardless of the tweet it belongs to. Then we find the language of this sentence (via Assumption 3) and tally up the totals for Hindi and English, for both corpora. Note that we are not saying that these results indicate that *users tend more to speak in Hindi than in English* (or vice versa), we are only saying that the conversation heuristics are henceforth.

Results shown in Table 3

**The skew is clearly visible between the total Hindi figures of the Dalit and Feminism corpora (64.43 percent - 22.43 percent respectively), indicating strongly that, in fact, the subject matter of a discourse affects the distributions of language within the discourse.**

**In more detail:**

The Dalit corpus, in fact, **is showing us a roughly 35 – 65 English-Hindi division**, which is counter to the general Twitter baseline of 70 – 30 English-Hindi[8].

What might be the reason for this? A brief look at the Dalit Corpus shows a good deal of charged, highly political discussion around national parties, policy and on-ground issues; it is not surprising that this conversation is embedded in the regional language, i.e. Hindi, especially since there is also campaigning and opinion-recruiting that is happening here. For example,
“priyankagandhi BhimArmyChief प्रियंका जी में केवल इतना जानना चाहूंगा कि आपकी सरकार ने आजतक हम दलितों के उद्धार के लिए क्या कदम उठाये है जात की राजनीति छोड़ अपने दम पर लड़ो”
(Priyanka Ji, all I want to know is what is your government doing on the behalf of us Dalits forget caste politics fight on your own strength)

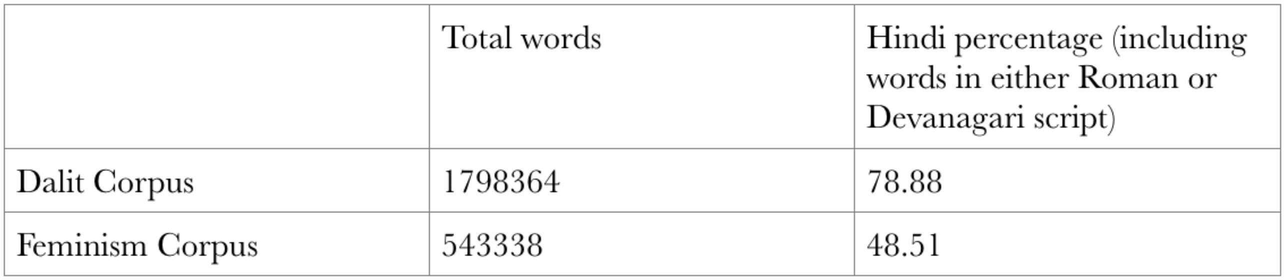
This entire genre of tweets, thanks to the growing Hindi trend/imposition on political propaganda by the current ruling party, is likely to occur in Hindi.

The feminism corpus, on the other hand, with its **78-22 English Hindi proportion**, slightly overshoots the Twitter English-Hindi baseline. Again, this aligns with our expectations: the conversation around feminism is still perceived to inhabit an elite upper-class arena, occupied majorly by the English-speaking educated class who will tend to choose English as the language of ideological discourse, perhaps slightly more than they would do given informal settings (as would be given a ‘normal’ discussion on Twitter). As we have also already discussed in (1), the key terms of discussion appear in English in Feminism Corpus far more than they do in Dalit Corpus – these results corroborate each other.

## VIII. Code-mixed fragments

We can also count the level, or granularity (Bali et al), at which code-mixing occurs in each corpus. A run-length

[8] Although there the unit is the tweet. Still, we can safely say that although our figures may not be exactly comparable, there is certainly marked divergence from the general English-Hindi baseline on Twitter



| | Total words | Hindi percentage (including words in either Roman or Devanagari script) |
|---|---|---|
| Dalit Corpus | 1798364 | 78.88 |
| Feminism Corpus | 543338 | 48.51 |

Figure 3: TABLE 3.

of 3 or more indicates that the fragment is larger than a tag (Bali et al 1617).

Most tags: ‘Hello’, ‘good morning’, ‘dekho toh’, ‘vaise bhi’, ‘kya batau’, ‘kuch nahi’, ‘bas’, ‘anyway’, ‘okay’, ‘cool’, ‘haan’, ‘haina’, ‘for example’ are not more than two words. We might have a few 3-word tags: ‘by the way’, for example, but we have chosen to keep the threshold run length as 3, as per convention. In any case (voila a 3-word tag), we have counted and displayed the number of sentences that exhibit CS fragments of all lengths from 1-20 in the more extensive results file.

We are avoiding naming lexical CS, because it is not within the scope of this project to be able to distinguish between CS and borrowing.[9] We have, however, dubbed 1-word CS fragments as ‘insertions’, and measured these anyway.

CS of higher granularity indicates greater ease between the two languages: for example, one need not be fluent in a language to be able to insert tags from it, whereas phrasal mixing would be a marker of greater fluency – in the context of our study: looking at the figures for phrasal CS will help us better understand the levels of co-immersion of the discussions in the languages concerned.

Results are recorded in Table 4

Here, we find that the figures match, indicating that people code-switch sentences with roughly equal frequency in both corpora. Phrasal CS, as we have specified it, occurs more in the Dalit corpus, indicating a slightly higher fluidity between Hindi and English in this conversation. This, again, is in line with all our previous results.

## IX. Code-mixed tweets

This is a user or tweet-based analysis. We want to say comparatively how many users code-mix on Twitter when participating in each discourse (as compared to the above, where we are calculating how much of the conversation is code-mixed).

We find, as we did for CS sentences above, that the figures match closely for both corpora, i.e. while people are making different broader linguistic choices, they switch between languages about equally frequently.

Results recorded in Table 5

The presupposition of our hypothesis was that the Dalit Corpus and Feminism Corpus would diverge in the Hindi-English proportions that they exhibit because of their content and sociolinguistic situations in the current Indian scenario.

**Accordingly, the above analyses demonstrate that relative Hindi presence as measured at various levels is consistently higher in the discourse on Dalit issues than it is in the discourse on women’s issues on Indian Twitter.**

## X. Conclusion

The results of this study indicate that there is a long way to go for the discourse of feminism in India to infiltrate to the grassroots level, even on social media. It also reflects the rising trend of politicism and propaganda in the Dalit discourse – the dominance of Hindi Devanagari, especially, rather than Romanized Hindi, which tends to be less formal, is significant. Feminism and the caste struggle are both being urgently mobilized in India, and while they might seem similar in being pro-equality movements, each is governed by its different context, history, and social perception, which is visible from their presence on social media. We are trying, through this study, to call for a great sensitivity to the sociolinguistic, political and historical dimensions of each movement when engaging with them on a national and local scale.

## XI. Works Cited

Bali, Kalika, Monojit Choudhury, Silvana Hartmann (2018). An Integrated Representation of Linguistic and Social Functions of Code-Switching. In LREC, 2018, pp 1615-1622.

Bali, K., Raafiya Begum and M. Choudhury. (2016). Functions of Code-Switching in Tweets: An Annotation Scheme and Some Initial Experiments. In LREC, 2016.

[9] This is also to reduce error due to the LID, which is occasionally eccentric and has uniformly labelled, for example, ‘patriarchy’ as Hindi (which is one kind of social commentary but still not very accurate.)