Accommodation of Conversational Code-Choice

Anshul Bawa **Monojit Choudhury** **Kalika Bali**
Microsoft Research
Bangalore, India
{t-anbaw,monojitc,kalikab}@microsoft.com

Abstract

Bilingual speakers often freely mix languages. However, in such bilingual conversations, are the language choices of the speakers coordinated? How much does one speaker's choice of language affect other speakers? In this paper, we formulate code-choice as a linguistic style, and show that speakers are indeed sensitive to and accommodating of each other's code-choice. We find that the salience or markedness of a language in context directly affects the degree of accommodation observed. More importantly, we discover that accommodation of code-choices persists over several conversational turns. We also propose an alternative interpretation of conversational accommodation as a retrieval problem, and show that the differences in accommodation characteristics of code-choices are based on their markedness in context.

1 Introduction

Code-switching (CS) refers to the fluid alternation between two or more languages within a conversation, and is a common feature of all multilingual societies. (Auer, 2013). Multilingual speakers are known to code-switch in spoken conversations for a variety of reasons, motivated by information-theoretic and cognitive principles, and also as a result of numerous social, communicative and pragmatic functions (Scotton and Ury, 1977; Söderberg Arnfast and Jørgensen, 2003; Gumperz, 1982).

Code-choice refers to a speaker's decision of which code to use in a given utterance, and in case of a CS utterance, to what extent the different codes are to be used. Depending on the sociolinguistic and conversational context, a speaker's code-choice may be unexpected and noticed by other speakers, and is likely to affect other speakers' subsequent code-choice. In other words, speakers may *accommodate* to each other's code-choice, positively or negatively (Genesee, 1982).

In this work, we propose a set of metrics to study the social accommodation of code-choice as a sociolinguistic style marker. We build upon the existing framework on accommodation by Danescu-Niculescu-Mizil et al. (2011) and adapt that for code-choice by introducing relevant features for code-choice. We then motivate and illustrate the effect of code markedness on the degree of accommodation - the more salient code is more strongly accommodated for. We further generalize the framework to also account for *delayed* accommodation, instead of only next-turn or *immediate* accommodation.

In addition, we introduce an alternative view of accommodation as a query-response task, and employ mean reciprocal rank, a well-understood metric from the domain of Information Retrieval, as a metric for latency of accommodation. We measure how quickly a style marker (code-choice in our case) introduced by a speaker is retrieved by the other speaker during the conversation. Our approach is developed for analyzing code-choice but is applicable to other dimensions of linguistic style as well (Tausczik and Pennebaker, 2010). This presents an alternative view of conversational style accommodation and offers a simple but effective way of measuring, characterizing and even predicting elements of conversational style.

We test this formulation on two CS conversational datasets - dialog scripts of bilingual Indian movies (in English and Hindi) and a transcription of real-world conversations between Spanish-English bilinguals in Florida, US. In both the corpora, we observe strong signals of interpersonal code-choice accommodation for the salient or marked code. We also observe that on average,

Proceedings of The Third Workshop on Computational Approaches to Code-Switching, pages 82–91
Melbourne, Australia, July 19, 2018.

the marked code is accommodated within the first three to four conversational turns, beyond which the effect of accommodation on code-choice decays gradually. Contextually-unmarked code is less strongly accommodated for, even when it occurs relatively infrequently within a conversation.

As far as we know, this is the first computational study of code-choice accommodation, and first work that introduces and formalizes the concept of delayed accommodation, that can be applied to other style dimensions as well.

The rest of the paper is organized as follows. We describe the background and related work in Section 2, which motivates the first formulation of code-choice accommodation in Section 3. We improve this by formulation by modifying the features in Section 4. We generalize the formulation to multiple turns and introduce the analogy to retrieval in Section 5, along with the results. We wrap up with a discussion in Section 6 that we conclude in Section 7.

2 Related Work

CS is employed by speakers to signal a common multilingual identity(Auer, 2005), and can be effectively used to reduce (or increase) the perceived social distance between the speakers (Camilleri, 1996). As a marker of *informality*, it has been shown to lower interpersonal distance (Myers-Scotton, 1995; Genesee, 1982).

Common structural patterns in CS as well as the choice to switch between languages have been the focus of many linguistic studies(Poplack, 1988)(Auer, 1995). As CS is typically used as a conversation strategy by bilinguals who are proficient in both languages (Auer, 2013), it is not surprising that certain pragmatic and socio-linguistic factors, such as formality of context (Fishman, 1970), age (Ervin-Tripp and Reyes, 2005), expression of emotion (Dewaele, 2010) and sentiment (Rudra et al., 2016), are found to signal language preference in CS conversations. A Twitter study of CS patterns across several geographies (Rijhwani et al., 2017), also suggests that there might be complex sociolinguistic reasons for code-choice. Thus, CS, and the choice of language or *code* in which one communicates during a multilingual conversation, could be considered a marker of *linguistic style*.

Communication accommodation theory (Giles et al., 1973; Giles, 2007) states that speakers shift their linguistic styles towards (or away from) each other in a conversation for social effect. In the CAT framework, the interlocutors' desire for 'social approval' results in an attempt to match each other's linguistic style. Accommodation has been studied for many markers of linguistic style like tense, negations, articles, prepositions, pronouns and sentiment (Taylor and Thomas, 2008; Niederhoffer and Pennebaker, 2002).

Since it is possible to convey the same semantic content while widely varying the extent of CS, we also consider code-choice as a linguistic style dimension. Therefore, we expect to observe accommodation in terms of code-choice in similar manner to that of variables for other linguistic styles. While there have been linguistic and small-scale studies (Sachdev and Giles, 2004; Bourhis, 2008; Bissoonauth and Offord, 2001; y Bourhis et al., 2007) that argue for prevalence of code-choice accommodation, there are no large-scale quantitative or computational studies that corroborate this and shed light on the various patterns of code-choice accommodation. Further, these studies rely on simple correlation-based measures.

The first computational study of linguistic style accommodation (Danescu-Niculescu-Mizil et al., 2011) shows that it is highly prevalent in Twitter conversations. They use binary features for the presence of various psychologically meaningful word categories as described by the LIWC method(Tausczik and Pennebaker, 2010) to identify stylistic variations in tweets. They then define a probabilistic framework that mathematically models style accommodation in terms of the likelihood of an addressee to respond in the same style as the speaker.

Though CS is similar enough to other kinds of linguistic style to allow analysis using the same framework, it also differs from them in being a strong sociological indicator of identity (Auer, 2005) and in not being processed nonconsciously (Levelt and Kelter, 1982). We demonstrate that a model that does not account for these crucial differences fails to capture the accommodative patterns of code-choice. Because of being processed consciously, code-choice also exhibits accommodation over several conversational turns, an effect which is not observed as strongly for other style dimensions (Danescu-Niculescu-Mizil et al., 2011). Long-term effects in accommodation have received very little attention, and have mostly

studied based on crude conversation-level correlation values (Niederhoffer and Pennebaker, 2002).

3 Accommodation of Code-Choice as Linguistic Style

As a first step, we adapt an existing framework (Danescu-Niculescu-Mizil et al., 2011) that quantifies *accommodation* of a given linguistic style. Any linguistic feature is said to exhibit accommodation if it is more likely to be expressed in response to a dialog that also expresses it, than otherwise. In other words, an accommodative feature in a dialog begets the same feature in the next dialog. We use the term 'dialog' or 'turn' to refer to a single spoken utterance or dialog within a conversation, and the term 'speaker' to refer to conversation participants. This framework thus restricts the definition of accommodation to only single-turn effects.

3.1 Measuring Accommodation

Mathematically, let F denote some binary feature over a dialog (we describe the features themselves in Section 3.2 below). F is said to exhibit accommodation if the likelihood of a user expressing F increases when F has been expressed in the previous dialog. We define the degree of accommodation as follows

$$Acm(F) = P\big(\delta_{d_i^F} | \delta_{d_{i-1}^F}\big) - P(\delta_{d_i^F}) \quad (1)$$

Here, dialog d_{i-1} immediately precedes dialog d_i, and δ_{d^F} is the *event* that the dialog d exhibits F. The first term can be thought of as the *reciprocity* over F. The second term is the fraction of dialogs in the corpus for which $F = 1$, which is also the empirical probability of observing F in a dialog d.

Instead of computing these likelihoods over the entire corpus, we could also compute them individually for each speaker, and doing so yields a fairer condition for accommodation. Different speakers can have widely different base likelihoods. This metric requires an average speaker to reciprocate more than their own (individual) baseline likelihood of expressing F, rather than simply more than the population baseline. Denoting the event that a dialog d is spoken by a speaker s as $\delta_{S(d)=s}$, we redefine accommodation as follows (E_s denotes an expectation over all speakers s)

$$\begin{aligned} Acm^*(F) &= E_s\big(Acm_s(F)\big) \\ &= E_s\Big(P\big(\delta_{d_i^F}\delta_{S(d)=s} | \delta_{d_{i-1}^F}\big) \\ &\quad - P\big(\delta_{d_i^F}\delta_{S(d)=s}\big)\Big) \end{aligned} \quad (2)$$

3.2 Measuring Code-Choice

Our general hypothesis is that code-choice is reciprocated in a bilingual conversation. To measure this, we introduce simple binary features for presence of each code, along the lines of the binary features in (Danescu-Niculescu-Mizil et al., 2011), with individual language expression substituting for the style dimensions. For each language L, we define a feature F_L indicating, for a dialog d, if the dialog contains words in the language L. The event that dialog d is at least partially in L, is denoted by $\delta_{d^{F_L}}$. In other words, $\delta_d^{F_L}$ is true if the language L is expressed in dialog d, and false otherwise.

3.3 Data

We employ two datasets of bilingual conversations, each in a different conversational context and a different pair of languages, to test the occurrence of code-choice accommodation. Table 1 reports the number of dialogs and words for the two datasets, and the fraction of words that are in English.

Dataset	Dialogs	Words	*%En*
Movies (*En-Hi*)	20.1K	240K	24.1
Bangor (*Es-En*)	18.5K	216K	62.9

Table 1: Conversational dataset overview

Hindi Movies

The data comprises of scripts of 32 Hindi movies released between 2012 and 2017. 17 of these scripts were collected by Pratapa and Choudhury (2017) from scripts posted online[1]. We collected 15 scripts of our own from a similar online source[2] and parsed them replicating the methodology of Pratapa and Choudhury (2017).

All the scripts have word-level language tags as created by the language identification system from (Gella et al., 2013). The language labels on manual inspection were found to have significant

[1]https: //moifightclub.com/category/scripts
[2]http://www.filmcompanion.in/category/fc-pro/scripts/

amount of noise, we corrected frequently observed errors with manual supervision.

Each dialog is assumed to be in response to the immediately preceding dialog within a scene. We restrict our analysis to dialogs that are between no more than two speakers, to avoid confounding effects of multi-party conversations on accommodation. This also filters out most dialogs in the scripts which are not conversational in nature.

Movie conversations, even though imagined, are designed to sound natural, and therefore, are suitable for studying style accommodation, as is argued in Danescu-Niculescu-Mizil and Lee (2011), and also multilingualism (Bleichenbacher, 2008) and code-choice (Vaish, 2011). It is true that movie dialogs promote stereotypes that may affect characters' expression of code-choice, however *accommodative effects* can still be expected to play out largely independent of such stereotypes. There have been several linguistic and quantitative studies on Hindi-English CS in Hindi movies (Parshad et al., 2016; Lösch, 2007; Pratapa and Choudhury, 2017).

Bangor Corpus

We use the Bangor Miami corpus[3] of word-level language labeled transcripts of spoken conversations between Spanish-English bilinguals in Florida, US. The original dataset contains 56 conversations, from which we selected 40 conversations that have non-trivial amount of English and Spanish, and sufficient dialogs from each speaker.

Figure 1 shows the fraction of Spanish used by a dyad of speakers in a sample conversation from this dataset (the complement fraction being English). Intuitively, we expect our metrics to capture how coordinated two speakers are.

3.4 Results

Table 2 shows the metrics from Section 3.1 computed over the features in Section 3.2 on the two datasets.

While these numbers do suggest that accommodative effects are present, they seem to be fairly weak. The rate of reciprocation is only slightly higher than the base rate, and in some cases the difference isn't statistically significant.

However, looking at individual differences in these values reveals an interesting observation. For each speaker s in the Movies dataset, we

[3]http://bangortalk.org.uk/speakers.php?c=miami

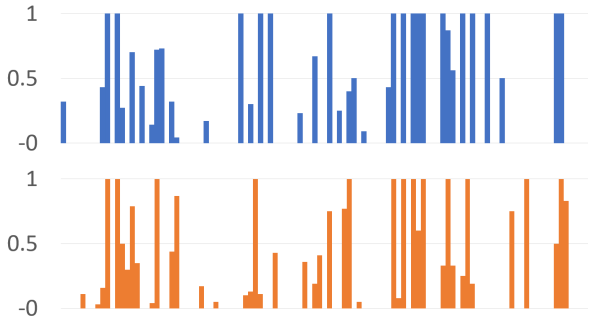


Figure 1: Fraction of Spanish over time in a conversation. The x-axis denotes consecutive dialog pairs, with dialog i above aligned with dialog $i+1$ below, so two aligned bars denote two consecutive dialogs.

Dataset	**Code** (L)	$Acm(F_L)$	$Acm^*(F_L)$
Bangor	En	0.06†	0.04
	Es	0.12	0.09
Movies	En	0.10	0.06
	Hi	0.02†	-0.02†

Table 2: Accommodation values for different codes. Values with a (†) are not significant. Significance for $Acm(F)$ is computed using Fisher's exact test, and significance for $Acm^*(F)$ is computed using one-tailed paired t-test.

plot in Figure 2, the rate of accommodation by s, $Acm_s(F)$, against the respective base rate $P(d_s^F)$, for $F \in \{F_{En}, F_{Hi}\}$.

Clearly, we see that a high base rate of expression corresponds to far less accommodation. In other words, the instances of code-choice that are uncommon and therefore unexpected within the conversational context are likely to be accommodated for. In a conversation that is predominantly in Hindi, a dialog uttered in Hindi carries little salience and doesn't stand out. This code-choice is unlikely to be registered as a communicative signal or a marked expression of any linguistic style, and therefore wouldn't elicit accommodation. English and Spanish are respectively less common in Movies and Bangor, and indeed their rates of accommodation are higher than the rates for the corresponding dominant languages.

Since the metrics in Section 3.1 compute likelihoods over all instances of code-choice irrespective of salience, the observed rates of accommodation are low. We borrow the notion of *markedness* of code-choice, as described in Myers-Scotton (2005), and incorporate it into our framework, as

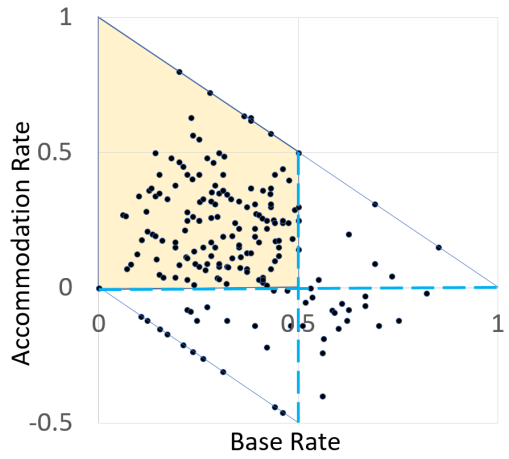


Figure 2: Variation of accommodation rate against base rate. Observed rate $(x + y)$ can vary between 0 and 1. The highlighted region denotes positive accommodation and a low base rate ($x < 0.5$ and $y > 0$). In contrast, all other regions, as demarcated by dashed lines, are sparser.

described in the next section.

4 Marked Code-Choice Features

4.1 Code Salience

As shown earlier, measuring accommodation makes sense only over marked instances of code-choice. Thus, for every conversation in our dataset, we identify the marked language, and measure accommodation only over that language. We choose a conversation as the unit for deciding if a code is marked because the set of speakers and the conversation context typically dictates code-choice in multilingual societies.

A language is considered marked if it is the non-dominant language - we keep the threshold of markedness at no more than 40% of total words in the entire conversation. We discard highly mixed conversations where none of the languages meets the threshold. This consideration also makes the calculation of accommodation more robust, as for a high fraction of incidence of a code, the effect of the previous turn would be harder to isolate.

4.2 Threshold of Occurrence

Another limitation of the formulation in Section 3 is that it doesn't incorporate the *extent* of presence of each code in a dialog. Consequently, even named entities, frequently borrowed words and frozen expressions from the marked language, would be considered as candidates for accommodation. The Bangor corpus came with named-entity tags, and in the Movies corpus we removed all character names from the dialogs, but we were not aware of any NER system for Hindi-English CS data that we could have used to remove other named entities. Ideally, we would like to exclude all such words from the triggers expected to elicit accommodation, as their usage isn't stylistically marked (Auer, 1999). The word-level language tags also have some amount of noise, and it is desirable to use features that are resilient to it.

Besides, it is possible that a relatively high incidence of marked code in a dialog is perceived as a stronger style marker, and is perhaps accommodated for more strongly than a lower incidence. We introduce a simple fraction-based thresholding that allows us to test the same.

For every dialog d, we define feature $F_{L,\tau}$ such that $d^{F_{L,\tau}} = 1$ if and only if (a) d is sufficiently long and (b) fraction of words of d in the marked language L is more than τ. We consider an utterance to be sufficiently long if it contains more than 4 words, as this is expected to filter out most frozen expressions and named entities that may be borrowed from one language to another. We show results for accommodation of F_τ for $\tau \in \{0, 0.2, 0.5\}$. While F_0 would capture presence of even one word in a marked code, $F_{0.2}$ represents a non-trivial occurrence and $F_{0.5}$ represents majority occurrence of the marked code in context.

5 Beyond Immediate Accommodation

The metrics in Section 3 and those in Danescu-Niculescu-Mizil et al. (2011) only consider the immediate next turn as a candidate for reciprocation. However, it is possible for accommodative effects to span a few conversation turns. Consider the following snippet from one of the conversations in Bangor (Spanish code is in bold and its translation is in italics).

In cases like this, the content of the conversation prevents a possibility of accommodating immediately, but the speaker *Sarah* still reciprocates *Paige*'s code-choice at the first instance possible. We can test if such cases of delayed accommodation are indeed common in the data, by extending our formulation to an arbitrary number of turns. We extend Equation (2) below, and Equation (1) can be extended analogously.

Paige	i wanna see them.
Sarah	pick. pick like (name) flowers or ...
Paige	**¡ay qué lindo está ese!**
	oh, how pretty that is!
	ok, enter the date.
	it will be ...
Sarah	may.
Paige	may. ninth?
Sarah	ninth.
Paige	two thousand and eight.
	and then you put what you want.
	(name) trip?
Sarah	**no te cabe.**
	it doesn't fit you.
	just (name).

5.1 Generalization of Immediate Accommodation

The baseline rate of a speaker s using a feature F across n (consecutive) turns is the likelihood that at least one the n turns expresses F, and is given by $1-(1-p_s)^n$, where p_s is simply $P(d_s^F)$. For a speaker s, the rate of n-turn accommodation is the increase in likelihood of occurrence of F in either of the n dialogs $d_{s,1}$ to $d_{s,n}$, conditioned on the event that the preceding dialog d_0 expresses F.

$$Acm_{n,s}(F) = P\Big(\bigvee_{i=1}^{n}\big(d_{s,i}^F\big)\Big|d_0^F\Big) - \big(1-(1-p_s)^n\big) \tag{3}$$

$$Acm_n^*(F) = E_s\big(Acm_{n,s}(F)\big) \tag{4}$$

When $n=1$, this resolves to Equation (2). Note that d_1 to d_n are the first n dialogs spoken by s immediately after the dialog d_0. As before, E_s denotes expected value over all speakers.

5.2 Accommodation as Retrieval

Responding to marked code-choice with marked code-choice can be thought of or reformulated as a retrieval task. For a speaker s, each instance of a dialog addressed to s with a feature F would be a *query* posed to s. The next n dialogs spoken by s would be the top-n retrieved *responses* to the query. We are interested in the retrieval of responses that also have feature F, so we call a response with feature F to be *relevant* response and *irrelevant* otherwise, in keeping with the standard terminology in information retrieval. We consider s to have retrieved a relevant response in n-turns if at least one of the first n responses is relevant.

When formulated this way, the *recall* of s, the probability of retrieving a relevant response, is precisely equal to the first term in Equation 3, the probability using F in responding to a dialog c^F. The second term in Equation 3 is the expected value of recall under the independence assumption, i.e., if s randomly introduces marked code at every turn with probability p_s. Therefore, a speaker is accommodative if their recall is higher than that of this random baseline.

A popular metric to evaluate retrieval systems is the mean reciprocal rank (MRR). The reciprocal rank of a query response is the multiplicative inverse of the rank of the first relevant response. The MRR of a system is simply the mean of the reciprocal ranks of all its responses. Since we expect the accommodative speaker to have a higher recall than the random baseline, we also expect the accommodative speaker to have a higher MRR, with the difference from baseline MRR being proportional to its accommodativeness.

Not only does this present an alternative view of accommodation and exposes well-studied formalisms and concepts from information retrieval, but the ability to capture speakers' styles as response characteristics also facilitates predictive conversational modelling.

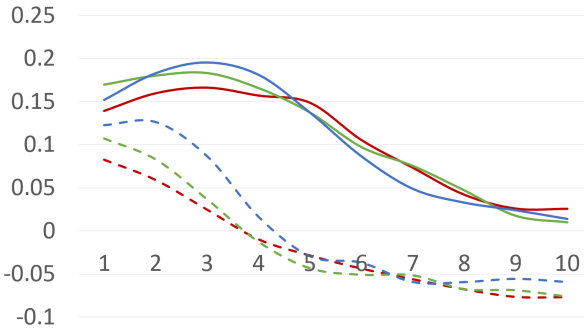
Mean reciprocal ranks for the random baselines can be computed analytically as follows. We first compute the expected reciprocal rank r for any given query as a function of the correctness probability p_s. For the first relevant response to be at rank i, all previous responses must be irrelevant. Since each response is relevant with a probability p_s, the probability of the i-th response being the first relevant response is given by :

$$P\big(r_{p_s}=\frac{1}{i}\big) = (1-p_s)^{i-1} * p_s \tag{5}$$

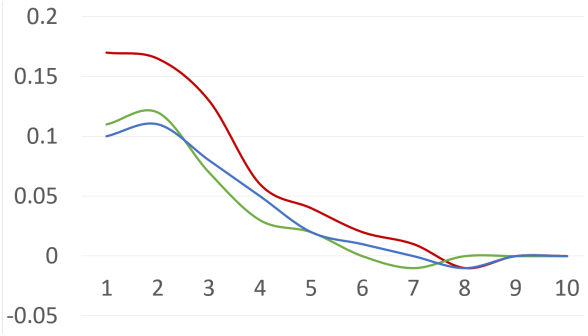
The baseline MRR of a speaker s, denoted by $Base_s$, is then the expected value of r, also as function of p_s :

$$\begin{aligned} Base_s &= E\big(r_{p_s}\big) \\ &= \sum_{i=1}^{\infty}(1-p_s)^{i-1} * p_s/i \\ &= -\frac{p_s}{1-p_s}\ln p_s \end{aligned} \tag{6}$$

The overall baseline MRR, $Base$ is then simply $E_s(Base_s)$. We compare the observed MRR on the data (denoted by Obs) with the expected MRR of the random baselines ($Base$), with their difference being indicative of the degree and immediacy of accommodation.



(a) Bangor; $L = Es$ for solid lines (significant for $n < 6$) and $L = En$ for dashed (significant for $n < 4$).



(b) Movies; $L = En$. Significant for $n < 4$.

Figure 3: Accommodation rates ($Acm^*_n(F_{L,\tau})$) versus n. Red, green and blue lines indicate τ = 0, 0.2 and 0.5 respectively. Accommodation of Hindi is not significant.

5.3 Results and Observations

Figure 3 shows the trends in $Acm^*_n(F_{L,\tau})$ for different values of n, L and τ. Significance scores are computed in the same way as for Table 2.

Table 3 shows the real and baseline MRR values for each corpus over different values of τ.

It is evident that accommodation of code-choice is a prevalent and robust phenomenon. The values of accommodation are consistently positive for all the different marked-code features, languages and datasets, and for low values of n.

In Table 3, the less common codes in each dataset, Es and En respectively, have a lower baseline while having comparable or even higher observed MRRs as their more common counterpart. This reiterates that accommodation is more pronounced for more marked codes.

From Figure 3, a higher fraction of marked code ($\tau = 0.5$) does not seem to elicit stronger accommodation than $\tau = 0$. However, it is important to note that the base rate for F_0 is much higher than that of $F_{0.5}$, so in relative terms, the latter exhibits

MRR	τ	**Bangor**		**Movies**	
		Es	En	En	Hi
Obs	0	0.48	0.52	0.67	0.62
$Base$		0.18	0.34	0.32	0.40
Obs	0.2	0.46	0.54	0.57	0.45
$Base$		0.15	0.26	0.25	0.35
Obs	0.5	0.46	0.49	0.54	0.4
$Base$		0.12	0.22	0.15	0.25

Table 3: Mean Reciprocal Ranks of the observed responses (Obs) and the random baseline ($Base$) for different features F_τ and different corpora.

a stronger tendency to accommodate (since the increase over respective base rate is identical). The difference between the retrieval characteristics for the different thresholds is more salient in Table 3 - higher thresholds correspond to a smaller average likelihood, and lower baseline MRRs. The difference between observed and baseline MRR does slightly increase with τ, making higher fraction of marked code somewhat more accommodated for.

In contrast to English, the accommodation for Hindi code-choice in conversations dominated by English is not significant. This suggests that Hindi code isn't marked even when it is the minority code in a scene, an inference that aligns with the claim from Myers-Scotton (2005) that Hindi is not marked in Hindi movies, even when it is the non-dominant language in context.

Hindi in Movies and English in Bangor have a lower strength of accommodation than their respective counterparts, even when measured over conversations where they are uncommon. Not only is accommodation stronger for Spanish, it also persists for more number of turns as compared to English. This suggests that the context of markedness is larger than the immediate conversation, and the being the dominant language of the corpus as a whole reduces markedness.

In most cases, accommodation is salient and significant even after a few turns. Delayed accommodation is as prevalent as immediate accommodation. And the likelihood of a given speaker reciprocating code-choice in kind, remains significant for several turns in a conversation.

6 Discussion

Accommodation is prevalent and robust, but not universal. While it is observed across conversations spanning different media and language

pairs, there is significant variation among speakers within a dataset. As many as 18% of the speakers exhibit what may be considered negative accommodation, or non-accommodation. Half of these do so with a value of $Acm_1^*(F_0)$ less than -0.10.

It is in fact known that accommodation or convergence is neither a universal nor a positive interpersonal strategy (Genesee and Bourhis, 1988; Giles et al., 1991; Burt, 1994). In-group/out-group identity as well as attitudes towards CS and the languages involved can cause negative accommodation as well as a negative perception of accommodation. Burt (1994) show that while convergence is largely viewed positively, some multilingual speakers may oppose it as either misplaced solidarity with an in-group, or a slur on the language capability of an interlocutor.

While we work under the assumption that code-choice is a style dimension, largely independent of content, it is in fact influenced by factors like topic (Sert, 2005) and sentiment (Rudra et al., 2016). These influences could either align or compete with the socially accommodative code-choice, and this explains several-turn accommodation - it is not always possible to accommodate immediately. The difference between code-choice and other linguistic style markers is also indicated by the poor results of Section 3, which naively applies the style accommodation framework to code-choice.

It is worth noting that the baselines throughout the paper assume that speakers do not adjust their overall rate of employing a particular code, in order to accommodate. This is in fact a fairly strict assumption. In fact, the same speaker typically has widely varying base rates in conversations with multiple other speakers. The extent of marked code to be used is itself often negotiated within a conversation, and adjusting one's base rate can be construed as accommodation, and harder to analyze. Nevertheless, this assumption gives us a strong and realistic baseline to judge the observations against.

One limitation of our formulation is that we do not look at individual words. Word or code saliency in context is actually more complex that just language saliency in current conversation. Some words are more marked than others, with borrowed words carrying very little salience. It would be nice to have more complex features, aware of the syntactic structure of dialogs. It would also be worthwhile to apply this formulation to study conversation-wide accommodation effects and convergence of code-choice at scale.

7 Conclusion

We demonstrate that code-choice is a marker of linguistic style, and when it is marked in context, it is interpersonally accommodated for. We extend the probabilistic formulation to multiple conversation turns, and show equivalence with a retrieval task, both facilitating better conversational analysis of code-choice in particular and style interactions in general.

In the future, we would like to use richer and linguistically motivated features for code-choice, including parts-of-speech, and indicators of borrowing across languages. Another generalization would be to also study LIWC words and markers of sociolinguistic style in this framework. Finally, longer-term accommodation effects, like convergence being succeeded by divergence, or topical effects on convergence, remain to be explored using a quantitative method like ours.

References

Peter Auer. 1995. The pragmatics of code-switching: a sequential approach. In Lesley Milroy and Pieter Muysken, editors, *One speaker, two languages*, pages 115–135. Cambridge University Press.

Peter Auer. 1999. From codeswitching via language mixing to fused lects: Toward a dynamic typology of bilingual speech. *International journal of bilingualism*, 3(4):309–332.

Peter Auer. 2005. A postscript: Code-switching and social identity. *Journal of pragmatics*, 37(3):403–410.

Peter Auer. 2013. *Code-switching in conversation: Language, interaction and identity*. Routledge.

Anu Bissoonauth and Malcolm Offord. 2001. Language use of mauritian adolescents in education. *Journal of Multilingual and Multicultural Development*, 22(5):381–400.

Lukas Bleichenbacher. 2008. *Multilingualism in the movies: Hollywood characters and their language choices*, volume 135. BoD–Books on Demand.

Richard y Bourhis, Shaha El-Geledi, and Itesh Sachdev. 2007. Language, ethnicity and intergroup relations. In *Language, discourse and social psychology*, pages 15–50. Springer.

Richard Y Bourhis. 2008. The english-speaking communities of quebec: Vitality, multiple identities and linguicism. *The Vitality of the English-Speaking*

Communities of Quebec: From Community Decline to Revival, Montréal, Centre détudes ethniques des universités montréalaises, Université de Montréal.

Susan Meredith Burt. 1994. Code choice in intercultural conversation. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 4(4):535–559.

Antoinette Camilleri. 1996. Language values and identities: Code switching in secondary classrooms in malta. *Linguistics and education*, 8(1):85–103.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754. ACM.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87. Association for Computational Linguistics.

Jean-Marc Dewaele. 2010. *Emotions in multiple languages*. Palgrave Macmillan, Basingstoke, UK.

Susan Ervin-Tripp and Iliana Reyes. 2005. Child codeswitching and adult content contrasts. *International Journal of Bilingualism*, 9(1):85–102.

J.A. Fishman. 1970. *Sociolinguistics: a brief introduction*. Newbury House language series. Newbury House.

Spandana Gella, Jatin Sharma, and Kalika Bali. 2013. Query word labeling and back transliteration for indian languages: Shared task system description. *FIRE Working Notes*, 3.

Fred Genesee. 1982. The social psychological significance of code switching in cross-cultural communication. *Journal of language and social psychology*, 1(1):1–27.

Fred Genesee and Richard Y Bourhis. 1988. Evaluative reactions to language choice strategies: The role of sociostructural factors. *Language & Communication*, 8(3-4):229–250.

Howard Giles. 2007. *Communication accommodation theory*. Wiley Online Library.

Howard Giles, Nikolas Coupland, and IUSTINE Coupland. 1991. 1. accommodation theory: Communication, context, and. *Contexts of accommodation: Developments in applied sociolinguistics*, 1.

Howard Giles, Donald M Taylor, and Richard Bourhis. 1973. Towards a theory of interpersonal accommodation through language: Some canadian data. *Language in society*, 2(2):177–192.

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

Willem JM Levelt and Stephanie Kelter. 1982. Surface form and memory in question answering. *Cognitive psychology*, 14(1):78–106.

Eva Lösch. 2007. The construction of social distance through code-switching: an exemplary analysis for popular indian cinema. *Department of Linguistics, Technical University of Chemnitz*.

Carol Myers-Scotton. 1995. *Social motivations for codeswitching: Evidence from Africa*. Oxford University Press.

Carol Myers-Scotton. 2005. *Multiple voices: An introduction to bilingualism*. Wiley-Blackwell.

Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Rana D. Parshad, Suman Bhowmick, Vineeta Chand, Nitu Kumari, and Neha Sinha. 2016. What is India speaking? Exploring the "Hinglish" invasion. *Physica A*, 449:375–389.

Shana Poplack. 1988. Contrasting patterns of codeswitching in two communities. *Codeswitching: Anthropological and sociolinguistic perspectives*, 215:44.

Adithya Pratapa and Monojit Choudhury. 2017. Quantitative characterization of code switching patterns in complex multi-party conversations: A case study on hindi movie scripts. In *Proceedings of the 14th International Conference on Natural Language Processing*, pages 75–84.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Sekhar Maddila. 2017. Estimating code-switching on Twitter with a novel generalized word-level language identification technique. In *Proceedings of the Annual Meeting of the ACL*.

Koustav Rudra, Shruti Rijhwani, Rafiya Begum, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. 2016. Understanding language preference for expression of opinion and sentiment: What do Hindi-English speakers do on Twitter? In *EMNLP*, pages 1131–1141.

Itesh Sachdev and Howard Giles. 2004. Bilingual accommodation. *The handbook of bilingualism*, pages 353–378.

Carol Myers Scotton and William Ury. 1977. Bilingual strategies: The social functions of code-switching. *International Journal of the sociology of language*, 1977(13):5–20.

Olcay Sert. 2005. The functions of code-switching in elt classrooms. *Online Submission*, 11(8).

Juni Söderberg Arnfast and J Normann Jørgensen. 2003. Code-switching as a communication, learning, and social negotiation strategy in first-year learners of danish. *International journal of applied linguistics*, 13(1):23–53.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Paul J Taylor and Sally Thomas. 2008. Linguistic style matching and negotiation outcome. *Negotiation and Conflict Management Research*, 1(3):263–281.

Viniti Vaish. 2011. Terrorism, nationalism and westernization: Code switching and identity in bollywood. *FM Hult, & KA King, K. A (Eds.). Educational linguistics in practice: Applying the local globally and the global locally*, pages 27–40.