User Perception of Code-Switching Dialog Systems

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

Abstract

Bilingual speakers often freely mix languages in conversation. Should dialog systems also be designed with an ability to code-switch, when interacting with multilingual users? In this paper, we explore this question based on a user-study on text-based bot-human conversations. Our results reveal three distinct classes of users with varying individual attitude towards code-switching (CS), and demonstrate the importance of a bot's CS fluency and its ability to reciprocate CS, in determining user preference. We also highlight some computational and sociolinguistic considerations that have implications for the design of multilingual dialog systems, and propose a strategy for dialog systems to navigate attitude estimation in mixed-language interactions.

1 Introduction

Code-switching (CS) is the fluid alteration between two or more languages within a conversation, and is common in most multilingual societies (Gumperz, 1982; Myers-Scotton, 1993). Multilingual speakers are known to code-switch in casual speech conversations for reasons motivated by its socio-pragmatic functions (Auer, 2013a; Begum et al., 2016; Auer, 1995), and driven by communicative and cognitive principles (Myslín and Levy, 2015; Scotton and Ury, 1977). As a marker of a shared multilingual identity (Auer, 2005), CS can make a conversation sound more natural and engaging, convey informality, and reduce perceived social distance between speakers (De Fina, 2007; Camilleri, 1996; Myers-Scotton, 1995).

Text-based conversational agents are now being developed in new languages (Shum et al., 2018). Although a large fraction of the world's population is multilingual (Ansaldo et al., 2008), nearly all conversational agents are still monolingual, which begs the following questions: Should dialog systems be designed to understand and respond in code-switched languages as well, or is it sufficient to simply have multiple monolingual agents. Given the communicative and social functions and roles of CS, can CS be an effective strategy for dialog systems? Can the appropriate use of CS by a dialog system improve its task-effectiveness or human judgment of its responses?

Further, would users perceive CS agents as being more natural or engaging, or do the social norms around human interactions not shape expectations for human-agent conversations, as is suggested by Ciechanowski et al. (2018)? For many agent applications, there are no obvious advantages of introducing CS ability, as only fluent bilinguals can code-switch and therefore, should be capable of conversing in either of the languages. Even if there are tangible improvements in judgment, would they be limited to certain types of users, or certain interaction-contexts? What would be the influence of reciprocity, and the quality of CS in the generated responses?

In this paper, we address the effect of CS usage on user perceptions through an in-depth user-study on human-bot conversations. Users observe snippets of human-bot conversations and are asked to compare several bot variants for naturalness and relative preference. We observe that users are frequently polarized on their judgments of CS, and that improved quality of generated CS also improves user judgments significantly.

We also observe that reciprocative use of CS is judged more favorably, indicating that the sociolinguistic theory of interpersonal accommodation (Bawa et al., 2018) also extends to CS in human-bot interactions. We carefully design the experiment and presentation to isolate the effect of CS from the effects of all other properties of the con-

Proc. of ICON-2018, Patiala, India. December 2018 ©2018 NLPAI, pages 166–174

versations and their dialogs.

Our multifaceted analysis reveals several interesting facts, such as:

1. Code-switching is used/perceived as a linguistic style marker, and therefore its accommodation is judged positively, even in human-bot conversations.
2. Irrespective of whether users code-switch themselves, their attitude towards a chatbot that code-switches can be extremely positive or negative.
3. Among users with a positive attitude, judgments of CS chatbots positively correlate with the naturalness of CS utterances, and with the accommodativeness of the bot in terms of its CS usage.

Thus, several important repercussions are borne out of this study in the context of dialog system design for multilingual societies. As far as we know, this is the first study to discuss the notion of CS in human-bot conversations.

We also discuss sociological and computational considerations affecting design choices of conversational agents, and briefly propose a novel strategy for navigating multilingual interactions and estimating user attitudes.

The rest of the paper is organized as follows- we motivate the questions of interest from related work in Section 2, followed by the experiment itself in elaborate detail in Section 3. We then discuss the implications of the study and its results in Section 4, and conclude in Section 5.

2 Motivation and Related Work

A wide range of differing attitudes towards CS has been well documented (Dewaele and Wei, 2014) and clearly indicate that CS is a style marker in multilingual conversations. Dewaele and Wei (2013) show that not only sociolinguistic factors like age, gender, education and language proficiency, but also personality types of speakers (levels of emotional stability, tolerance to ambiguity, cognitive empathy and neuroticism) affect their attitude towards CS. While we show CS to be similar to other dimensions of linguistic style (Tausczik and Pennebaker, 2010) in its cohesive and accommodative characteristics in human conversations, it also differs from them in being a strong sociological indicator of identity (Auer, 2005). Because of this sociological dimension, users may have different attitudes towards CS, and these attitudes may vary with users' demographic profiles. We delineate such effects in the study.

A computational study of style accommodation (Danescu-Niculescu-Mizil et al., 2011) shows that style-accommodation is highly prevalent and exhibits great complexity in Twitter conversations.

Language interaction and socio-pragmatic utility of code-switching in multilingual societies is very well studied in linguistics (Scotton and Ury, 1977; Fishman, 1970; Ervin-Tripp and Reyes, 2005; Dewaele, 2010; Rudra et al., 2016). Due to the prevalence and naturalness of CS in human-human conversations, we argue that a CS agent can build better rapport with its user by connecting to their common multilingual identity. Further, 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), have been found to function as socio-pragmatic signals for language preference in CS conversations. Style convergence in a conversation signals warmth and reduced inter-personal distance (Myers-Scotton, 1995; Blom et al., 2000), and Bawa et al. (2018) have shown that choice of language (or code) exhibits interpersonal convergence or accommodation in human conversations. Therefore, users could expected to follow similar patterns of conversation with an agent, and could find responses that follow these patterns to be more natural, which makes a case for both a CS understanding and generation ability in conversational agents.

However, it is unclear how much of the human-human conversation norms reflect in, and have the same pragmatic effect in, human-agent conversations. While there is evidence that humans exhibit a "chameleon effect" when engaged in a social-interaction (Ward and Litman, 2007; Reitter et al., 2011), there is limited evidence of any convergence in human-agent interactions (Branigan et al., 2010).

We do not know of any human-agent study that explores the effects of multilingual usage and code-switching. Hill et al. (2015) show that there are significant linguistic differences between human-human and human-agent conversations, which might be due to users' adaptation to the limitations of the technology (Arif and Stuerzlinger, 2012; Vinciarelli et al., 2015). These observations suggest that in a multilingual society, users may effortlessly adapt to monolingual agents, just as they would to monolingual speakers. Our in-depth user study extends some of these findings to code-switching as a dimension of linguistic style, by measuring how code-choice and code-choice accommodation by conversational agents are perceived by different kinds of human users.

Though, as shown in another user study (Thies et al., 2017) on preferences over bot personalities that even a single user can have different preferences based on what they are trying to achieve, we might argue that depending on the context of usage of an agent, user might prefer a CS agent (for instance for personal assistant or chit-chat) or a monolingual agent (for a particular goal, like booking flight tickets, or finding scientific articles), and while multilingual users can easily adapt to a monolingual agent, a CS agent could still be perceived as more empathetic and engaging, providing a better overall user experience.

Yet another confounding factor is the wide range of differing attitudes towards CS that exist at the level of an individual or community of speakers (Dewaele and Wei, 2014). Dewaele and Wei (2013) show that not only sociolinguistic factors like age, gender, education and language proficiency, but also personality types of speakers (levels of emotional stability, tolerance to ambiguity, cognitive empathy and neuroticism) affect their attitude towards CS. Thus, the preference towards a CS agent might widely vary across individuals, communities and multilingual geographies.

Given the wide differences in the perception of CS, perception of chatbots, and the fact that insights from human-human conversations cannot be trivially applied to human-agent conversations, we cannot a priori comment on the usefulness of CS agents, and perform a user study to gain insights on these questions.

3 User Study on Human-Bot Conversations

The goal of the study is to quantify the causal effects of the following on the user judgments of agents:

1. Presence of CS in agent and human dialogs
2. Expressed attitude of users towards CS, as revealed and inferred from user comments
3. Naturalness and reciprocal nature of CS
4. Demographic profile of the users

Users are shown a series of human-bot conversation snippets in pairs, and asked to rate the agents (bots) on conversational skill and relative preference within the pair. The users are also asked to comment on the aspects of the conversations that they notice, the differences between the two agents' conversational skill, and the reasons for their stated preferences. These snippets average about 15 dialogs each in length, and vary in the presence of CS in dialogs by either dialog participant.

3.1 Experiment Design

There are four variants, or 'conditions' of each presented conversation : No CS by either participant ($None$), CS used by agent only ($Agent$), CS by human only ($Human$) and CS by both participants ($Both$). This is done to isolate the effect of *style* (here CS) from the effect of *content* of the conversations.

In each trial, a user is shown two human-agent conversation snippets, each with a different agent. Users are asked to rate each agent on perceived conversational skill on a 7-point Likert scale, ranging from 'Extremely Bad' to 'Extremely Good', and to compare the two agents by assigning relative preference between the two, again on a 7-point Likert scale, with 'Strong preference' for either agent as the extremes and 'No preference' as the mean.

In a pilot version, users were shown conversation pairs that differed only exactly in CS usage and were identical otherwise. This led to starkly negative judgments of CS, as CS usage stood out and was therefore judged critically. Since code-switching is a non-conscious process for many fluent bilinguals (Heredia and Altarriba, 2001), explicitly asking users to judge CS in this way is likely to distort judgments because of observer bias (Poulton, 1975).

Therefore, to mask the variables of interest, we conduct a single-blind study, and allow the human-agent conversations to differ not just in CS usage but in various other respects such as conversational topic, lengths of utterances, expressed sentiment and conveyed personality. We do not control for any variation across conversations, stylistic or otherwise, and treat all differences between conversations (except code-switching) as confounding covariates, or noise.

3.1.1 Presentation Order

Users are randomly divided into four groups; each user sees 4 pairs of conversations to be compared (so 8 conversations per user), where one was always in the $None$ condition, while the other varies across all four conditions mentioned above (including the $None$-$None$ pair, which provided the baseline judgments that we later use to measure the ratings for other conditions against). We employ a presentation order similar to the case-crossover study design (Lombardi, 2010), in which different randomly-assigned groups see different permutations of conversations and CS variants, as shown in Table 1.

Each group sees the same conversations (and in the same order) as other groups, with the corresponding conversations across the groups differing exactly in the CS conditions applied. Additionally, as the order of presentation of the CS variants varies across the groups, it takes care of ordering biases like fatigue and carryover (Strack, 1992).

	Conversation Pair			
Group	**1**	**2**	**3**	**4**
1	$None$	$Agent$	$Human$	$Both$
2	$Agent$	$Human$	$Both$	$None$
3	$Human$	$Both$	$None$	$Agent$
4	$Both$	$None$	$Agent$	$Human$

Table 1: Presentation order of the CS variants

3.1.2 Conversation Generation Methodology

The conversation snippets are extracted from real conversations between one of the authors and an online chatbot, Ruuh[1] that were conducted as a part of this study. We chose to use real conversations with a chatbot rather than artificially constructed chat so that the data reflects the nature and technological constraints of human-bot conversations.

Eight such conversation snippets were collected that span everyday topics such as hobbies, travel, relationships and work. A fluent English-Hindi bilingual was consulted to generate code-switched versions of the dialogs ($Agent$, $Human$ and $Both$ variants). In keeping with the findings on markedness of conversational English-Hindi CS from Bawa et al. (2018), the generated conversations have Hindi as the marked language and moderately high values of sentence-level fraction of Hindi words.

The box below shows one such conversation. The full list of questions and all conversations are provided in Supplementary Material.

Human	Do you play any musical instruments?
Bot	Nope, just listen to a lot of music I guess.
Human	*Main* piano play *karna sikh raha hoon* ⟨ I'm learning to play the piano ⟩
Bot	Piano *bajaanaa* is like a zen mood *mere liye*. ⟨ Playing the piano is like a zen mood for me ⟩
Human	Indeed, it is very relaxing to play the piano after a long day at work
Bot	Does learning how to play the piano help?
Human	*Isse* focus *karne me* help *hota hai*, and to develop a taste in music ⟨ It helps you to focus, and to develop a taste in music ⟩
Bot	Music *to ek* language *hai yaar*. You were born to speak it. ⟨ But music is a language dude. You were born to speak it. ⟩
Human	Yes, but some nuances need to be learnt.
Bot	No I do not think music can be learnt.

Sample conversation. Hindi-code is in *italics*.

3.2 Analysis

A total of 68 users participated in our study, all fluent English-Hindi bilinguals. A user typically took 15-20 minutes to answer all questions.

3.2.1 Demographic Profile

In addition to the judgment scores from users, we collected basic demographic information about each user- their age, native language and other languages known, places where they've lived in for at least three years, and highest level of education attained. Interestingly, none of these were found to be correlated with either attitude classes or with judgments. (exact values of these correlations are reported in Supplementary Material).

3.2.2 Comparing Judgment Ratings

The 7-point skill ratings are normalized to zero-mean and unit-variance for each user, as we are interested only in the *relative* ratings given to the CS variants by each user. The 7-point preference ratings over conversation pairs are similarly normalized to unit-variance.

Let $\textsc{Skill}_{V,C}$ denote this normalized skill rating for a conversation C presented with the CS variant V, as judged by all users who see this variant. Let $C1$ and $C2$ be the two conversations presented in some pair, with $C1$ presented

[1] https://www.facebook.com/Ruuh/

with CS variant V and $C2$ with the base variant ($None$). The corresponding relative skill rating of $C1$ over $C2$ is then defined as $\textsc{Skill}_{V,C1,C2} = \textsc{Skill}_{V,C1} - \textsc{Skill}_{None,C2}$.

We adjust these ratings against the base difference between the conversations $C1$ and $C2$ (differences that can be attributed to everything except the code-switching) by getting $\textsc{Skill}'_{V,C1,C2} = \textsc{Skill}_{V,C1,C2} - \textsc{Skill}_{None,C1,C2}$. We are able to do this because every conversation pair is also shown once to *some* group without any CS in either conversation ($None$ condition in Table 1). The overall skill rating of condition $V \in \{Agent, Human, Both\}$ is then just the average over all conversation pairs, $\textsc{Skill}_V = \mathbb{E}_{C1,C2}(\textsc{Skill}'_{V,C1,C2})$. Let $\textsc{Pref}_{V,C1,C2}$ denote the normalized preference rating of $C1$ over $C2$, as judged by all users who see this pair. We analogously derive the overall preference rating $\textsc{Pref}_V$.

3.2.3 Inferring Attitude Classes

We look at the text comments provided by users and classify users based on if they explicitly mention language mixing (or any paraphrasing of it) in one of the differences that they notice between the conversation pairs, and if they express any sentiment towards it. This gives us three types of users, differing in their expressed attitude towards CS:

Pos : CS is noticed; positive attitude or preference expressed. (39% users)

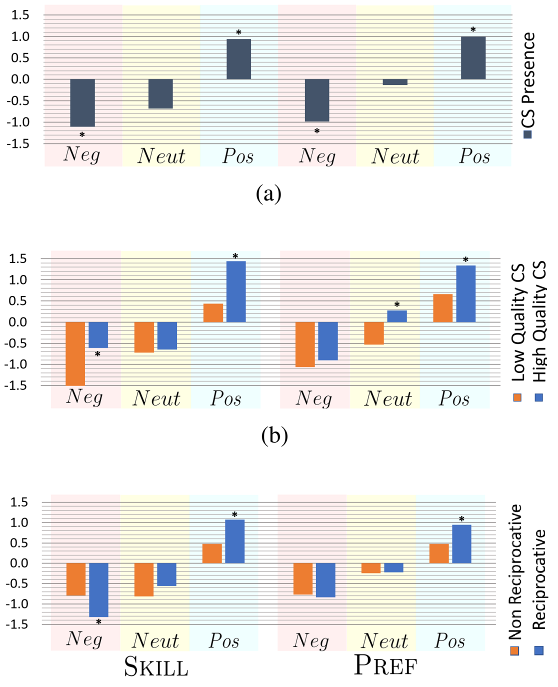
Neg : CS is noticed; negative attitude or dispreference expressed. (29% users)

$Neut$: CS is not mentioned or no sentiment can be discerned from comments. (32% users)

We do not observe any significant trends in values of $\textsc{Skill}_V$ and $\textsc{Pref}_V$ across the population as a whole, but the distribution of these values does suggest some clustering of user ratings. Indeed, when we condition the ratings $\textsc{Skill}_V$ and $\textsc{Pref}_V$ on the attitude class $A \in \{Pos, Neg, Neut\}$ of the users providing them, denoted by $\textsc{Skill}_V^A$ and $\textsc{Pref}_V^A$, we see significant differences.

3.2.4 Labeling CS Quality

We are interested in quantifying the effect of CS quality on user judgments. We had two independent annotators rate all conversations and variants on a 5-point scale for CS fluency and naturalness. We then binarize this judgment and label each conversation as having 'high-quality CS' or 'low-quality CS'. This divides the 4 conversations into two classes of two conversations each. Restricting $\textsc{Skill}_V$ only to conversations with CS quality $Q \in \{High, Low\}$, gives $\textsc{Skill}_V^Q$.



(a)

(b)

(c)

Figure 1: Skill and preference ratings across various conditions and user groups. (*) denotes significant difference from zero or between the pair.

3.3 Observations

Figure 1a shows $\textsc{Skill}^A_{Bot+HumanBot}$ and $\textsc{Pref}^A_{Bot+HumanBot}$ for all A. They capture the effect of presence of CS in bot's dialogs. Figure 1b shows $\textsc{Skill}^{Q,A}_{Bot+HumanBot}$ and $\textsc{Pref}^{Q,A}_{Bot+HumanBot}$ for all Q and A, and they show the effect of quality of CS of the bot's dialogs. Figure 1c) shows $\textsc{Skill}^A_{HumanBot}$ against $\textsc{Skill}^A_{Bot+Human}$ and $\textsc{Pref}^A_{HumanBot}$ against $\textsc{Pref}^A_{Bot+Human}$ for all A. This brings out the effect of accommodative CS on SKILL and PREF judgments.

3.3.1 Presence and Quality of CS

The effects of CS on these normalized judgments of skill (SKILL) and relative preference (PREF) for the users in the three attitude classes are seen in Figure 1. The effect of presence and quality of CS in agent dialogs are shown in Figure 1a and 1b respectively. For the attitude class Neg, all judgments of CS (irrespective of quality) are consistently negative, which is not surprising, as users who have a dispreference for the phenomenon of CS itself are unlikely to have a notion of quality.

SKILL judgments by $Neut$ users are similar to those in Neg, but PREF judgments are sensitive to CS quality. We speculate that their dispreference can at least partially be addressed by improving CS generation by bots, or by aligning it better with known user CS patterns.

3.3.2 Reciprocity of CS

Figure 1c compares the ratings when only one of the parties code-switch ($Agent$ and $Human$) to when both code-switch in a reciprocative manner ($Both$). Pos users have a strong preference towards reciprocative CS, which is in line with reciprocity observed in human conversations (Bawa et al., 2018). This suggests that users in this class judge conversational agents on similar parameters as they would judge human interlocutors. Users in Pos not only rate CS highly, but are also sensitive to both quality and accommodativeness of bot's CS, with accommodative CS perceived as much more skillful than anti-accommodative CS.

3.4 Results

To summarize, the study points to two primary takeaways: (1) it is important to *know* or otherwise *identify* the "user's attitude" towards an agent that code-switches, as introducing CS has diametrically opposite effects on users with different attitudes, and (2) quality of CS responses is important to all users, and might also influence their attitude towards a CS-agent in the long run.

Overall, it is suggested from the study that good-quality accommodative CS significantly improves judgment for a large fraction of users whose general attitude towards CS has otherwise been identified.

Demographic factors are all poorly correlated with judgments of SKILL and PREF. Demographic variables also fail to predict attitude classes, with all classes having a similar spread of all demographic variables. See Supplementary Material for all such correlations.

4 Discussion

The findings of the study have multiple implications for the design of CS conversational agents, and their strategy for making CS decisions.

4.1 Attitude Estimation

The ability to infer a user's attitude towards CS seems to be the single-most important determinant of the success of any CS strategy by a bot, as a bot that can infer the attitude class can make an informed decision on whether or not to adopt CS. Users' attitude towards CS depends on various social and psychological variables (Dewaele and Wei, 2013). Since attitude estimation is crucial, it should be incorporated into the design of an agent. Individual disposition could be predicted at the demographic level, though we found no correlation between users' attitude and their demographics, suggesting that demography-based inference of attitude is unreliable.

Furthermore, it is not necessary that a person's attitude towards CS in human conversations matches that in a human-bot conversation. For instance, in our study, users commented: "I didn't like the way he was switching languages. That felt very forced." and "...is trying too hard to sound natural".

It seems straightforward to just ask the user about their CS preferences, but as CS choices could be nonconscious or spontaneous decisions like other aspects of linguistic style (Levelt and Kelter, 1982), stated preferences are unlikely to be as reliable as preferences revealed from observing users in-conversation (Levitt and List, 2007). Furthermore, the latter estimates would also reveal a user's individual style of CS and extent of CS.

4.2 Nudging as a Conversational Strategy

Such probing of CS preferences while conversing needs to be a balancing act. Without any adaptation, the agent is likely to stick to a suboptimal default. On the other hand, aggressive probing in the form of arbitrary CS decisions will immediately distract and annoy users.

We propose *nudging* as a strategy to navigate this tradeoff - the agent slightly deviates from its default response (in terms of CS) and measures if the change is reciprocated, either immediately or over a few turns, to implicitly estimate user preference. In principle, these changes can also be measured using other evaluation criteria (Liu et al., 2016). Nudging has been studied for human-agent interactions in other non-conversational contexts (Sadigh et al., 2016) and with the aim of influencing user behavior over digital interfaces (Weinmann et al., 2016). We propose nudging as an active exploration strategy to reveal user preferences, over which longer term policies can then be studied. In our conversational context, 'nudging' would mean that the bot introduces instances of marked code-choice gradually in increasing amounts of markedness, while being sensitive to its measured effects on the user. These effects, in turn, could be inferred from their replies, from factors such as user's code-choice and/or trends in expressed sentiment.

4.3 CS Quality

We also see from the study that the perceived quality of generated CS matters, as bilinguals easily spot inaccuracies or unnatural CS patterns. Some of their responses within the user-study, to poor-quality CS, are: "answers in the style of speaker, all right, but the hinglish is unnatural", "... The place to change the language is a bit unnatural according to me" and "I don't like the way codemixing was used". Perceived CS quality affects judgments regardless of attitude.

Unless conversational systems can consistently generate high-quality CS, they may not be very well-received. CS quality itself depends on multiple factors. The first and most basic factor is syntactic soundness of mixed sentences. Joshi (1985) is the only work we know of that computationally generates grammatical CS sentences. The second determinant of perceived quality is the statistical likelihood of the CS pattern, which depends on the strength of the underlying language model. While there have been several proposed language models for CS (Adel et al., 2015; Ying and Fung, 2014), none of them have been evaluated against human judgments.

Quality judgments could be learned from natural CS data over Twitter, and there is initial work on making artificially generated CS similar to natural CS (Pratapa et al., 2018). Perceived CS quality, however, goes beyond the surface form of a sentence and is influenced by social and pragmatic functions in context (Begum et al., 2016).

While a better codification of naturalness judgments is needed, we speculate that an initial and safe strategy for an agent to explore nudging would be to introduce simpler CS constructions, like tags, frozen expressions (Poplack, 1988), or frequently observed discourse markers (Auer, 2013b). Monolingual responses from an existing conversational agent could be modified with such constructs in simple, even rule-based ways, to get corresponding CS versions.

5 Conclusion and Future Work

Our user study shows that proficient and accommodative CS in bots improves their perceived skill and preference for users who have a positive attitude towards CS by bots. In conclusion, we have argued that conversational agents need to discover and adapt to user CS preferences in order to gain relevance in multilingual contexts.

We propose nudging as a way to infer, on-the-fly, the users' attitude towards CS agents. In future work we would like to explore the utility of this strategy through a Wizard-of-Oz study, where the bot (wizard) can adaptively change the code-choice based on user's responses. Such strategies could also be automatically inferred from analysis of human-bot conversations in an inverse reinforcement learning framework as formulated in (Sadigh et al., 2016). Further, unlike our user study, a Wizard-of-Oz study will allow users to actively participate in the conversation and therefore provide better judgments.

References

Heike Adel, Ngoc Thang Vu, Katrin Kirchhoff, Dominic Telaar, and Tanja Schultz. 2015. Syntactic and semantic features for code-switching factored language models. *IEEE Transactions on Audio, Speech, and Language Processing*, 23(3):431–440.

Ana Inés Ansaldo, Karine Marcotte, Lilian Scherer, and Gaelle Raboyeau. 2008. Language therapy and bilingual aphasia: Clinical implications of psycholinguistic and neuroimaging research. *Journal of Neurolinguistics*, 21(6):539–557.

Ahmed Sabbir Arif and Wolfgang Stuerzlinger. 2012. How do users adapt to a faulty system. In *CHI'12 Workshop on Designing and Evaluating Text Entry Methods*, pages 11–14.

Peter Auer. 1995. The pragmatics of code-switching: A sequential approach. *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, pages 115–135.

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

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

Peter Auer. 2013b. The'why'and'how'questions in the analysis of conversational code-switching. In *Code-switching in Conversation*, pages 164–187. Routledge.

Anshul Bawa, Monojit Choudhury, and Kalika Bali. 2018. Accommodation of conversational code choice. *Third Workshop on Computational Approaches to Linguistic Code-switching, ACL 2018*.

Rafiya Begum, Kalika Bali, Monojit Choudhury, Koustav Rudra, and Niloy Ganguly. 2016. Functions of code-switching in tweets: An annotation framework and some initial experiments. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Jan-Petter Blom, John J Gumperz, et al. 2000. Social meaning in linguistic structure: Code-switching in norway. *The bilingualism reader*, pages 111–136.

Holly P Branigan, Martin J Pickering, Jamie Pearson, and Janet F McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics*, 42(9):2355–2368.

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

Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor. 2018. In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems*.

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.

Anna De Fina. 2007. Code-switching and the construction of ethnic identity in a community of practice. *Language in society*, 36(3):371–392.

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

Jean-Marc Dewaele and Li Wei. 2013. Is multilingualism linked to a higher tolerance of ambiguity? *Bilingualism: Language and Cognition*, 16(1):231–240.

Jean-Marc Dewaele and Li Wei. 2014. Attitudes towards code-switching among adult mono-and multilingual language users. *Journal of Multilingual and Multicultural Development*, 35(3):235–251.

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.

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

Roberto R Heredia and Jeanette Altarriba. 2001. Bilingual language mixing: Why do bilinguals code-switch? *Current Directions in Psychological Science*, 10(5):164–168.

Jennifer Hill, W Randolph Ford, and Ingrid G Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, 49:245–250.

A. K. Joshi. 1985. Processing of Sentences with Intrasentential Code Switching. In D. R. Dowty, L. Karttunen, and A. M. Zwicky, editors, *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, pages 190–205. Cambridge University Press, Cambridge.

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

Steven D Levitt and John A List. 2007. What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic perspectives*, 21(2):153–174.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

David A Lombardi. 2010. The case-crossover study: a novel design in evaluating transient fatigue as a risk factor for road traffic accidents. *Sleep*, 33(3):283–284.

Carol Myers-Scotton. 1993. *Dueling Languages: Grammatical Structure in Code-Switching*. Claredon, Oxford.

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

Mark Myslín and Roger Levy. 2015. Code-switching and predictability of meaning in discourse. *Language*, 91(4):871–905.

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

EC Poulton. 1975. Observer bias. *Applied ergonomics*, 6(1):3–8.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *ACL*.

David Reitter, Frank Keller, and Johanna D Moore. 2011. A computational cognitive model of syntactic priming. *Cognitive science*, 35(4):587–637.

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.

Dorsa Sadigh, S Shankar Sastry, Sanjit A Seshia, and Anca Dragan. 2016. Information gathering actions over human internal state. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 66–73. IEEE.

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.

Heung-Yeung Shum, Xiaodong He, and Di Li. 2018. From eliza to xiaoice: Challenges and opportunities with social chatbots. *arXiv preprint arXiv:1801.01957*.

Fritz Strack. 1992. order effects in survey research: Activation and information functions of preceding questions. In *Context effects in social and psychological research*, pages 23–34. Springer.

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.

Indrani Medhi Thies, Nandita Menon, Sneha Magapu, Manisha Subramony, and Jacki Oneill. 2017. How do you want your chatbot? an exploratory wizard-of-oz study with young, urban indians. In *IFIP Conference on Human-Computer Interaction*, pages 441–459. Springer.

Alessandro Vinciarelli, Anna Esposito, Elisabeth André, Francesca Bonin, Mohamed Chetouani, Jeffrey F Cohn, Marco Cristani, Ferdinand Fuhrmann, Elmer Gilmartin, Zakia Hammal, et al. 2015. Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions. *Cognitive Computation*, 7(4):397–413.

Arthur Ward and Diane Litman. 2007. Dialog convergence and learning. *Frontiers in Artificial Intelligence and Applications*, 158:262.

Markus Weinmann, Christoph Schneider, and Jan vom Brocke. 2016. Digital nudging. *Business & Information Systems Engineering*, 58(6):433–436.

Li Ying and Pascale Fung. 2014. Language modeling with functional head constraint for code switching speech recognition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 907–916.