Sociolinguistic Factors of Mandarin-English Codeswitching: Language Attitudes, Age, and Other Factors Used for Computational Modeling

Irene Yi
Yale University

Sociolinguistic Factors of Mandarin-English Codeswitching: Language Attitudes, Age, and Other Factors Used for Computational Modeling

Abstract

This paper explores the sociolinguistic predictors of Mandarin-English codeswitching, and also tests such patterns against current syntactic constraints of codeswitching. By doing so, I demonstrate the value of incorporating sociolinguistic factors as predictors into computational models of codeswitching, explored in a companion paper (Yi 2022). The study presented here draws from novel data collected from 12 Mandarin-English bilingual speakers from Grand Rapids, Michigan. These speakers come from two generations, correlated with their age and immigration history. Speakers participated in sociolinguistic interviews that were designed to elicit codeswitching in narrative-style responses on a variety of topics, including family, school, and culture. Participants also answered metalinguistic questions about their own language practices and attitudes and completed a written Language History Questionnaire (LHQ) (Li et al. 2020), which asked for self-evaluations of language habits (proficiency, immersion, and dominance in the two languages). LHQ responses were then quantified into "scores" that served as sociolinguistic predictors for the companion paper (Yi 2022). Patterns found in this novel Mandarin-English data frequently, and potentially systematically, violate many of the currently proposed syntactic constraints on codeswitching (which mainly come from research on Spanish-English bilinguals), implying that the constraints may not be universal, and that new avenues should be considered for understanding the morphosyntax of bilingual codeswitching.

Sociolinguistic Factors of Mandarin-English Codeswitching: Language Attitudes, Age, and Other Factors Used for Computational Modeling

Irene Yi*

1 Introduction

The twofold goal of this paper is as follows: firstly, to investigate nuanced differences in linguistic identities and language attitudes of Chinese American bilingual speakers based on sociolinguistic factors (e.g. age, education level, language immersion, etc.); and secondly, to compare newly collected codeswitched data against previously proposed syntactic constraints on codeswitching. These subparts revolve around the central question of how and why the Chinese-American community represented in this study codeswitches.

The novel data for this study were collected from the Chinese-American community in Grand Rapids, Michigan. Codeswitched speech was elicited from families who identify as part of the Chinese Association of West Michigan (CAWM) along two age groups: Younger (20–30 years old) and Older (over 45). These correlated with the participants' immigrant generation: whether they immigrated to the United States as an adult or whether they are children of immigrants, born in the US. The collected speech described in this paper was then transcribed and tokenized into a corpus of sentences, with the presence or absence of a codeswitch in each sentence marked. This, in turn, set the base for a companion paper (Yi 2022) that develops a Classification and Regression Tree (CART) model trained on transcribed Mandarin-English bilingual codeswitched speech.

2 Background

2.1 Syntactic Constraints on Codeswitching

Bilingual and multilingual speakers often "switch" back and forth between their languages, either between or within sentences. This phenomenon, known as codeswitching (CS), is a common practice of speakers of more than one language. There are many proposed constraints and theories surrounding CS, and this paper will focus on three: the equivalence constraint (Poplack 1978), the free morpheme constraint (Poplack 1978), and Functional Head Constraint (Belazi, Rubin, and Toribio 1994). These will be laid out below in the section discussing violations of these constraints from my data.

I lay out these three constraints to show that the theories on CS are not universally accepted. This is likely due to the fact that what is allowed in CS varies greatly cross-linguistically, and a large proportion of the CS literature (e.g. Poplack 1978, Solorio and Liu 2008) is focused on only English-Spanish CS (or CS between English and another Romance language). As with any proposed linguistic universal, the eurocentric approach often gives a limited template that is hard-pressed to fit a typologically different language, or even just a language that was not represented when creating such theories.

Instead of forcing data that do not fit certain constraints (e.g. FHC) into these frameworks (Huddlestone 2002) by trying to explain one's non-English-Spanish-CS data from the view of English or Spanish syntax, this paper will forefront the data as it is and argue that CS data does not have to fit one or all of these constraints to be deemed as grammatical or "correct" CS patterns. Further, as this paper focuses in part on the bilingual linguistic identity of individuals who identify as bilingual in Mandarin and English, any CS uttered by the participants in the data collection process will be seen as grammatical, unless otherwise noted by the participants themselves. Belazi, Rubin and Toribio

* I would like to thank Professor Gašper Beguš and Professor Isaac Bleaman (both UC Berkeley), my advisors and supervisors for this project. Thank you to the participants of my study, as well as the pilot study participants who helped me hone my methodology. Additionally, I would like to thank Professor Claire Bowern (Yale University) for everything she has taught me this year. Finally, a big thank you to my friends and family who have supported and loved me throughout this work.

1994 asserted that violations of FHC could be used to gauge whether or not a speaker was actually bilingual, or whether there was an imbalance in bilingualism and fluency of both languages. While speakers can use different frequencies of one language over the other in CS (or in their speech in general), this does not invalidate their bilingualism; certainly, violations of FHC do not invalidate their bilingualism. Rather, this paper centers sociolinguistic factors such as age (and by extension, immigrant generation) to account for the differences in representation of either Mandarin or English in their CS. In this way, the syntax of CS is used more as a way to analyze the data that does exist, instead of putting value judgements (and externally imposed grammaticality judgements) on the utterances.

2.2 Sociolinguistic Identity Construction Through Language

A sociolinguistic approach of identity and language patterns, then, rather than a purely syntactic one, would be much more informative for our purposes. The study in this paper focuses on a community of Chinese Americans in Grand Rapids, Michigan.

Grand Rapids, Michigan is 2.4% "Asian" (US Census 2019), with no specification of what proportion of that statistic is specifically Chinese or Chinese American. From personal experience and the experience of study participants, Grand Rapids is fairly homogeneous in terms of most sectors of public life being English-speaking spheres. Additionally, there were not many public school resources in the early 2000s for students whose first language was not English. Because of this, the children of Chinese immigrants often felt pressure to assimilate or adapt into monolingual American society. This often manifested in insecurities about speaking Mandarin in public settings, or neglecting to learn and maintain Mandarin altogether. Zheng 2018 outlines a similar experience of Chinese Americans in the city of Troy, Michigan. Troy, at 19% Asian and 5% having self-identified as being of Chinese descent, has one of the largest Chinese American populations in Michigan; in fact, it is deemed "the Asian city of southeast Michigan" (Zheng 2018). Despite a higher percentage of the population identifying as Asian, Troy shares Grand Rapids' sentiment of feeling insecure about speaking and maintaining Mandarin.

While regional language habits (particularly in research into the dialectology of English as spoken by white Americans) take up much attention in sociolinguistics research, the language habits of minority groups—particularly Asian American or Chinese American communities—have been studied to a much lesser extent. Zheng 2018 notes in particular that, of the existing research on Chinese Americans, most is focused on coastal Chinese American communities (i.e. New York City Chinatown, San Francisco Chinatown, etc.) rather than Midwestern Chinese Americans. This gap is significant because Midwestern communities are more homogeneously white compared to coastal regions, which affects the cultural and linguistic experience of any minority group living in the Midwest. Namely, Midwestern Chinese communities face more pressure to assimilate, both culturally and linguistically, into white communities, which inevitably influences these communities' language habits and attitudes (Zheng 2018).

Codeswitching was historically (and to a degree, still is) misinterpreted as a speaker's lack of fluency in one or more languages, though Poplack (1980:581) disproves this, stating that "codeswitching, rather than representing debasement of linguistic skill, is actually a sensitive indicator of bilingual ability". However, CS is often still stigmatized as an indicator that the speaker lacks bilingual fluency, even today. In fact, one participant in the current study assumes that CS is an indicator of being bad at either Mandarin, English, or both. In a qualitative study by Yim and Clément (2019) on Cantonese-English codeswitching in Toronto, participants noted very mixed attitudes towards their own CS habits. Even in a very diverse community such as Toronto, where there is a vibrant Chinatown and Chinese community, outsiders' negative evaluations of CS affected speaker attitudes. Additionally, in both a pilot study of the experiment design as well as in an actual participant interview, it was noted by the speakers that CS was looked down upon in an academic setting, such as a classroom at school. Further, some expressed that it is seen as unprofessional to codeswitch in any environment that wasn't in the home, with very close friends, or both. Other participants said they felt some degree of shame in codeswitching, as it made them feel like they appeared as not proficient in their two languages, even if they were fully bilingual.

A central aim of this paper is to validate the bilingualism, language habits, and lived experiences of the individuals in this specific Chinese-American community. A second goal of this paper is to

provide literature representing the codeswitching habits of this community in an effort to destigmatize CS; if even a single person who relates to the experiences of this community can feel less ashamed of their own CS habits, this paper will have been successful.

3 Methodology

3.1 Data Collection

Novel data were collected through sociolinguistic interviews that elicited Mandarin-English codeswitched speech. These interviews were conducted remotely over Zoom due to the COVID-19 pandemic, and the procedure was approved by the Institutional Review Board (IRB) before any of the data collection process began. In order to elicit codeswitched speech, each interview included the same list of questions that were themselves asked using codeswitching between Mandarin and English. By using fairly balanced codeswitching between the two languages in the way the questions were written, participants would not be primed from the manner of question-asking to lean towards answering fully or mostly in one language or the other.

It is crucial for sociolinguistic analysis to collect data that are as naturalistic as possible, so this danger of biasing participants needed further investigation to minimize any avoidable influence in the experiment design. This question-language bias was explored in preliminary pilot studies, where data was not actually collected, but different experiment designs (i.e. the list of questions and what language they were asked in) were tested. The pilot studies were not run with the actual participants of the study, but rather Mandarin-English bilinguals who volunteered to help me hone my sociolinguistic interview process. The purpose of these pilot studies was to see what the effects of the interviewer's speech (i.e. me) were on the participants' responses. Pilot study volunteers noted that questions asked in codeswitching-style itself felt the most like natural conversation, which was a goal for the experiment design. A key advantage of being a bilingual researcher in this experiment design was that I was able to elicit as naturalistic of CS data (by asking questions using CS and speaking with CS) as possible.

The final form of the interview included 22 questions that were designed to elicit narrative-style speech from participants. These were asked in codeswitched speech, and I used codeswitching in my speech between questions during the interview Zoom calls as well (such as when I responded to their answers to each question). Zoom calls were recorded with participants' consent, as well as IRB approval. Each Zoom call began with a warm-up activity where participants were asked to name simple cartoon images in both Mandarin and English, to get them comfortable with using both languages on the call. However, these parts of the Zoom recordings were not transcribed for data analysis, as this portion of the call was not representative of the interview or naturalistic speech. Lastly, the Zoom calls ended with 17 metalinguistic questions where I asked participants about their own language habits and their views on codeswitching. The answers to these questions were also not transcribed as part of the data because they were often one-word answers. The topics of conversation included everyday life, family (in the United States), family (in China), friends, school, work, food, plans for the future, media consumption, and dreams participants had while sleeping.

Data were collected from a total of 12 participants, six of whom were in the older age category (>= 45 years old), and six of whom were in the younger one (20–30 years old). The older group of speakers ranged in age from 45 years old to 63 years old, and the younger age group of speakers were all in their early 20s. The line dividing the age group was drawn to correlate with their immigration generation. The six participants in the older age group all immigrated to the US as adults, while the six participants in the younger age group were born in the US as children of immigrants. All twelve participants are bilingual in Mandarin and English, and they are all familiar with the practice of codeswitching. Most participants did not know the name "codeswitching" for this process; rather, they called it "Chinglish." All participants were part of the Chinese Association of West Michigan (CAWM), located in Grand Rapids, Michigan. CAWM is a cultural organization that provides a community to the Chinese population of Grand Rapids and West Michigan as a whole. CAWM provides many cultural services, including food festivals, holiday celebrations, and the Grand Rapids Chinese Language School (a weekend language school).

Each person participated in their own Zoom call interview, making a total of 12 recorded interviews of codeswitched speech. Each call lasted about 40 minutes on average, though the first few

minutes of the interview (the warm-up activity about naming images) and the final ten or so minutes (metalinguistic questions) were not transcribed as part of the data analysis. Additionally, I did not include transcriptions of my own speech in the data analysis, as most of what I said were the pre-written interview questions. Thus, about 20–30 minutes of speech were manually transcribed for each participant. The total speech time transcribed was 4 hours, 39 minutes, and 12 seconds.

Participants additionally answered metalinguistic questions about their own language practices and attitudes and completed a written Language History Questionnaire (LHQ) (Li et al. 2020), which asked for self-evaluations of language habits (proficiency, immersion, and dominance in the two languages). LHQ responses were then quantified into "scores" that served as sociolinguistic predictors for the current study's companion paper (Yi 2022).

4 Results

4.1 Language Behavior and Attitudes

There were 1340 sentences transcribed in the final (cleaned) dataset. Out of the 1340 sentences, 309 contained at least one instance of CS. 640 sentences were spoken by the older age group and 700 were from the younger age group. There was a total of 16,285 words, 7064 of which were English words and 9221 of which were in Chinese. Of the 9221 Chinese words, 7082 were spoken by the older age group and 2139 were spoken by the younger generation. Of the 7064 English words, 1045 were uttered by the older generation and 6019 came from the younger generation.

Li et al. 2020 lay out the calculations behind LHQ "scores", and Yi 2022 describes the choices to use these scores for this particular paper in more detail. The scores used here were L1 proficiency, L2 proficiency, L1 immersion, L2 immersion, L1 dominance, L2 dominance, L2 to L1 dominance ratio, and the multilingual language diversity (MLD) score. Scores for all participants are in Table 1.

Participant	L1 Prof.	L2 Prof.	L1 Imm.	L2 Imm.	L1 Dom.	L2 Dom.	L2:L1 Dom.	MLD
1	0.99	0.84	0.72	0.56	0.59	0.63	1.08	1
2	0.99	0.86	0.99	0.68	0.68	0.5	0.73	1.31
3	0.76	0.99	0.54	0.54	0.47	0.66	1.4	0.98
4	1	0.6	0.54	0.54	0.66	0.31	0.48	1.21
5	0.99	0.71	0.97	0.73	0.59	0.54	0.92	1
6	0.79	0.63	0.98	0.5	0.46	0.37	0.81	0.99
7	0.74	0.64	0.98	0.55	0.11	0.38	3.53	0.76
8	1	0.63	0.67	0.65	0.66	0.3	0.45	1.37
9	0.84	1	0.54	0.54	0.44	0.87	1.97	1.38
10	1	0.84	0.78	0.76	0.62	0.5	0.8	0.99
11	0.71	0.99	0.76	0.5	0.57	0.58	1.01	1
12	1	0.75	0.5	0.5	0.73	0.38	0.52	0.92

Table 1: LHQ scores of speakers.

Disregarding where codeswitches are and how many sentences were spoken, Table 2 shows the raw distribution of words in each language by each age group, as well as the corresponding chi-square test for these data.

	Chinese utterances (words)	English utterances (words)	Percentage of Chinese utterances per age group	Percentage of English utterances per age group
Older Age Group	7082	1045	0.871	0.129
Younger Age Group	2139	6019	0.262	0.738

Table 2: Distribution of language utterances across age groups: Chi-square statistic is 6152.0636, and the p-value is <.00001, meaning the difference in language use by age group is statistically significant at $p < .05$.

As Table 2 shows, the older age group uses more Chinese utterances, while the younger age group uses more English utterances. This could be a reflection of the immigration generation, as the older age group immigrated to the US as adults who did not become immersed in an English-speaking environment until later in their lives. They all have, however, worked in the US for at least two decades in English-speaking environments. Younger generation speakers were born in the US and became immersed in English-speaking environments at school growing up, and may be used to using more English even when speaking with a bilingual Chinese American interlocutor (i.e. me).

The metalinguistic questions of interviews revealed interesting insights about the language attitudes and habits of speakers. Every single one of the twelve participants said that they speak in codeswitched language ("Chinglish") with their immediate family members who lived in the US. In other words, the younger generation used CS when speaking with their parents and siblings, if they had any. The older generation used CS when speaking with their children (and siblings if they had any who lived in the US), but used only Mandarin when speaking to their parents (who, if alive still, mostly live in China, Hong Kong, or Taiwan). Everyone stated that they mostly used English in work or school contexts, regardless of age group, though the older generation mentioned that they would codeswitch when talking about work to their other Chinese American friends of similar ages. The younger generation mainly said that they would codeswitch with other non-family Chinese Americans when mainly speaking about family, food, or memories in China. When talking about media consumption of media forms in Chinese, both age groups tended to use more Chinese utterances than when talking about a different topic of conversation, though the younger generation still used more English utterances than the older generation. When talking about media consumption of English media, the same patterning happened, but with English instead of Mandarin.

For both age groups, using Mandarin reminded speakers of their family, heritage, and culture. All participants mentioned that speaking in Mandarin helped them index parts of their identity to interlocutors, and that they would often use Mandarin in Chinese-dominated public social situations (e.g. at a restaurant in Chinatown in the US) to communicate a sense of solidarity in identity with, at times, total strangers (e.g. a random Chinese person or Chinese American in Chinatown). When asked about codeswitching specifically, every participant said that they have codeswitched before and would codeswitch with at least one social circle in their lives (e.g. family), but the attitudes towards codeswitching differed across speakers. For the younger generation, which was above mentioned to correlate somewhat with an undergraduate education level, speakers saw it either as a normal part of their linguistic habits or as a slightly "shameful" habit due to the misconception that CS is a sign that one is not fully proficient in one or both of their languages.

Younger speakers expressed that older or more educated speakers would look down on "speaking Chinglish" as a sign that the younger speaker did not know a certain word in Mandarin, or was simply not very fluent in Mandarin. However, I knew these younger speakers are fluent in Mandarin, and often whichever word or phrase they codeswitch into English for, they will also use the Mandarin word for that same concept later in their speech. In other words, they are not codeswitching due to a lack of knowledge for a certain word, but likely for whichever language they accessed first when thinking about that word or phrase. However, because Grand Rapids has a small Chinese community compared to the other communities represented in Michigan, younger generations expressed a sense of wanting to keep their Chinese identity, culture, or heritage. Thus, when older speakers misperceive CS as a sign that a young speaker does not have a good enough grasp of

Mandarin, younger speakers sometimes feel insecure, as though this marked them losing their cultural identity. Because of this, many of the already fluently bilingual speakers in the younger generation are now taking college-level Mandarin courses to practice speaking as well as reading and writing. While they are bilingual speakers with a strong multicultural identity either way, the Mandarin classes they take help boost their sense of identity and sense of security in their culture and language. However, in these Mandarin classes, younger speakers have reported that habits of codeswitching are discouraged in speaking (again, as it is seen as unacademic to codeswitch), and even more so in writing, thus furthering the perception of CS as something that should not be done in academic or professional spheres.

Older generation speakers reported that they do not feel as insecure about their own codeswitching habits, since there is no doubt in their mind that they have a solid Chinese cultural identity, but they still look down on the younger generation for practicing the same language habits of codeswitching. As mentioned above, older generation speakers will codeswitch with their friends of the same age quite frequently in social settings, and while their children may codeswitch the same amount when speaking to parents, one interlocutor relationship where CS is present is stigmatized more. Thus, CS is a very nuanced part of speaker identities from many different aspects: who is speaking (including the age and generation of the speaker), who they are speaking to, what topic they are speaking about, the speaker's relationship with their own cultural identity, and the public perception of this speaker's cultural identity by others.

4.2 Violations of Syntactic Constraints

Part of the reason CS is so stigmatized in academic settings is likely due to the idea that there is a right or wrong way to use language. Because Mandarin-English CS is understudied and proposed syntactic constraints on CS (as mentioned in the literature review) are largely based on languages that are not Mandarin (e.g. Spanish-English CS), the phenomenon of Mandarin-English CS is not fully understood by linguists, thus making Mandarin-English CS have constructions seen as ungrammatical by syntactic constraints' standards. Further, a native speaker who codeswitches between Mandarin and English may have only been exposed to the ways that CS is misperceived as linguistic shortcoming by community members, so the stigmatization that they know of CS (combined with even linguistic or syntactic constraints that deem certain constructions as "ungrammatical") might make it an even less favorable language habit to practice. As this paper is not focused on the syntax of Mandarin, English, or CS, I will not attempt to completely revise proposed constraints on CS, but it would do my participants a disservice to not present the data that contradicts and violates well-known constraints. If nothing else, we can at least understand parts of constraint-violating Mandarin-English CS that show up in speech of speakers across age groups, education levels, and more. (1), (2), and (3) below come from my data, and they violate the equivalence constraint (Poplack 1978), the Functional Head Constraint (Belazi et al. 1994), and the free morpheme constraint (Poplack 1978), respectively.

(1)	我	有	applied	for	graduate	school
	1SG	PST	apply.PST	PREP	ADJ	NOUN

'I have applied for graduate school'

The equivalence constraint (Poplack 1978) states that intra-sentential CS can only occur at points of shared syntactic boundaries between the two languages. This means that the syntax of the two languages must line up so that the separate pieces of different languages fit together. In (1), this is violated because of the redundancy of the past tense marker. 'PST' is marked on both the English word applied (in the form of the past tense suffix *-ed*) and the Mandarin word 有. The constraint states that we would expect there to be only one T head, but tense is doubly marked in this example. In Mandarin, there are different morphemes that can be used fairly interchangeably that mark past tense, and 有 'PST' is one of them. Normally, if a speaker used 有, they would not use any other tense marking. Therefore, in both Mandarin and English, marking tense on two parts in the sentence is redundant and does not follow the syntactic boundaries of either language. However, this is a perfectly grammatical and understandable codeswitched sentence that was uttered by a native speaker of both languages. Other speakers used similar redundant constructions in their speech as

well. (2) shows a violation of Functional Head Constraint (Belazi, Rubin, and Toribio 1994):

(2)	I	think	that	我	会	再	apply	for	jobs
	1SG	VERB	COMP	1SG	FUT	again	apply.PRES	PREP	job.PL

'I think that I will apply for jobs again (in the future)'

This is a clear and straightforward example. Functional Head Constraint (FHC) asserts that a codeswitch cannot happen between a functional head and its selected complement. In this case, the functional head is the Complement Head (C) that and its selected complement, the Tense Phrase (TP) of 我会再 apply for jobs 'I will apply for jobs again.' The C is in English and switches partially into Mandarin for the TP. This construction of a codeswitch happening between a C' and a TP was extremely common in the sentences uttered by all of my speakers.

(3) shows a violation of the free morpheme constraint (Poplack 1978), which says that codeswitches cannot happen between a lexical head and a bound morpheme. In (3), then, we see the violation of a codeswitch happening between the Mandarin lexical head 读 'read' and the English bound past tense morpheme *-ed* 'PST':

(3)	我	haven't	读ed	书	in	a	while
	1SG	have.PST.NEG	read.PST	book.PL	PREP	DET	NOUN

'I haven't read books in a while'

The violation of the constraint is quite clear with the lexical head and its bound morpheme being in different languages. Though this is not a full syntactic analysis of Mandarin-English CS, it is clear that there needs to be further investigation in this direction if linguists are to propose syntactic constraints on codeswitching.

4.3 Quantitative Analyses

When only looking at the number of sentences that contain the presence of at least one CS point, there were 309 sentences that contained CS (as mentioned above), with 137 of those being from the older age group and the remaining 172 from the younger age group. Just over half of all CS-containing sentences came from the younger age group (55.66% of all CS-containing sentences), while 44.34% came from the older age group. A narrow majority of non-CS-containing sentences came from the younger age group as well (51.21%). The total number of sentences with no CS at all was 1031 sentences, with 503 from the older group and 528 from the younger group. However, these percentages may simply reflect the fact that the younger age group accounts for 52.24% of the total sentences spoken in the data. This potentially misleading comparison is seen in Figure 1:

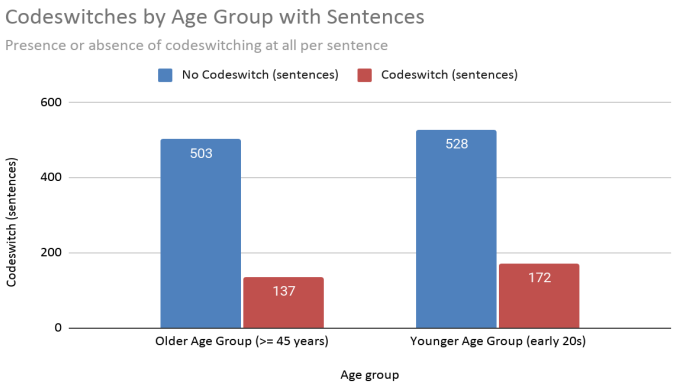

Figure 1: Comparing the number of sentences that do or do not contain CS across age.

Figure 2 shows the drastic difference across age groups between the number of sentences where codeswitching was simply present and the actual number of CS occurrences by each group. There were a total of 162 total instances of CS by the older group over 137 sentences, averaging 1.18 instances of CS per sentence where it was present. The younger age group, however, had 432 total instances of CS across 172 sentences where it was present, which makes an average of 2.51 instances of CS per sentence.

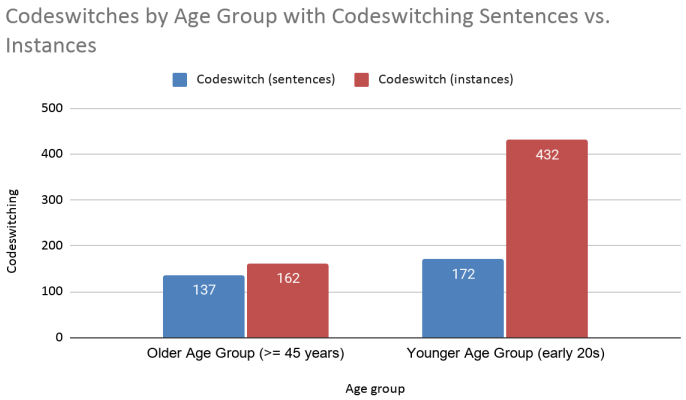

Figure 2: Codeswitches by age group with sentences vs. instances.

However, I still wanted to consider how much more the younger generation actually codeswitches than the older generation. To do this, I compared the number of CS instances per age group with the total words each group uttered. As mentioned above, 16,285 words is the grand total uttered by both groups combined. The older generation had 8127 words total, where 7965 were not instances of CS points and 162 were (about 2.0% of total utterances). The younger generation had 8158 words total, where 7726 were not instances of CS points 432 were (about 5.3% of total utterances). While these percentages may seem small, it is quite unlikely that every single word that did not have a CS point would be a potential CS instance anyway; rather, comparing the frequency between the two age groups is insightful. Table 3 below presents the data:

Utterances of CS per Age Group	Non-CS utterances (instances)	CS utterances (instances)	Total utterances	Percentage of CS utterances compared to total utterances
Older Age Group	7965	162	8127	0.020
Younger Age Group	7726	432	8158	0.053

Table 3: Utterances of CS per age group: Chi-square statistic is 126.3091, and the p-value is <.00001, meaning the proportions of CS to non-CS across the two age groups is statistically significant at $p < .05$.

5 Conclusion and Discussion

This paper investigated the nuance of codeswitching habits and how they index identity, and compared novel data against proposed syntactic constraints of CS. Such patterns violating constraints are uttered by speakers of all ages, education levels, and LHQ scores, which implicates that it may be a feature of Mandarin-English CS that should be syntactically accounted for rather than a simple sociolinguistic variant in the syntax of codeswitched speech.

In the future, sociolinguistic analyses of codeswitching should be researched to a much greater and much more nuanced extent, especially as it can work in conjunction with the already existing syntactic and semantic ways to model language in computational research on such phenomena (see further Yi 2022).

On a non-technological side, more research should be done into the phenomenon of codeswitching, specifically in Mandarin-English CS or CS involving languages whose morphosyntactic structure is vastly different than English. This can not only help our linguistic understanding of such a process (so that we do not have constraints that are constantly violated in languages that were not taken into account when proposing the constraints in the first place), but it can also work to destigmatize the language habit of codeswitching and empower many communities and generations of speakers who codeswitch. This, in turn, will help speakers construct their linguistic identity in the context of their cultural identity and heritage without feeling the shame that current young speakers who codeswitch between Mandarin and English do.

References

Belazi, Hedi, Edward Rubin, and Toribio, Almeida Jacqueline. 1994. Code switching and X-Bar Theory: The functional head constraint. *Linguistic Inquiry* 25:221–237.

Huddlestone, Kate. 2002. Grammatical constraints on intrasentential code switching: Evidence from English-Afrikaans code switching. *Stellenbosch Papers in Linguistics PLUS* 31:91–113.

Li, Ping, Fan Zhang, Anya Yu, and Xiaowei Zhao. 2020. Language History Questionnaire (LHQ3): An enhanced tool for assessing multilingual experience. *Bilingualism: Language and Cognition* 23:938–944.

Poplack, Shana. 1978. Syntactic structure and social function of code-switching. In *Latino Language and Communicative Behaviour*, ed. R. Duran. New York: Ablex Publishing Corp.

Poplack, Shana. 1980. Sometimes I'll start a sentence in spanish y termino en espanol: Toward a typology of code-switching. *Linguistics* 18:581–618.

Solorio, Thamar, and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, ed. M. Lapata and H.T. Ng, 973–981.

U.S. Census Bureau. 2019. Grand Rapids, Michigan.

Yi, Irene. 2022. Sociolinguistically-aware computational models of Mandarin-English codeswitching. In *Proceedings of the Linguistic Society of America 7.1*, ed. P. Farrell, 5247.

Yim, Odilia, and Richard Clément. 2019. "You're a Juksing": Examining Cantonese–English code-switching as an index of identity. *Journal of Language and Social Psychology* 38:479–495.

Zheng, Mingzhe. 2018. You have to learn to adapt: A sociolinguistic study of Chinese Americans in the "Asian city" of southeast Michigan. Doctoral dissertation, Michigan State University.

Department of Linguistics
Yale University
New Haven, CT 06511
ireneyi@berkeley.edu