Jee haan, I'd like both, por favor: Elicitation of a Code-Switched Corpus of Hindi–English and Spanish–English Human–Machine Dialog

Vikram Ramanarayanan & David Suendermann-Oeft

Educational Testing Service R&D
90 New Montgomery Street, Suite 1500, San Francisco, CA
<vramanarayanan,suendermann-oeft>@ets.org

Abstract

We present a database of code-switched conversational human–machine dialog in English–Hindi and English–Spanish. We leveraged HALEF, an open-source standards-compliant cloud-based dialog system to capture audio and video of bilingual crowd workers as they interacted with the system. We designed conversational items with *intra*-sentential code-switched machine prompts, and examine its efficacy in eliciting code-switched speech in a total of over 700 dialogs. We analyze various characteristics of the code-switched corpus and discuss some considerations that should be taken into account while collecting and processing such data. Such a database can be leveraged for a wide range of potential applications, including automated processing, recognition and understanding of code-switched speech and language learning applications for new language learners.

Index Terms: code switching, human-computer interaction, dialog systems

1. Introduction

Code-switching refers to multilingual speakers' alternating use of two or more languages or language varieties within the context of a single conversation or discourse in a manner consistent with the syntax and phonology of each variety [1]. Linguists in particular have extensively studied this phenomenon (see for example [2, 3, 4, 5, 6]). The literature identifies *inter*-sentential (alternation between sentences, also called extra-sentential) and *intra*-sentential (within sentences, can also include intra-word) switching as two of the primary types of code-switching observed in bilinguals [7, 8].

Theoretical linguistic interest aside, an important motivating factor for studying and developing tools to elicit and process code-switched language comes from the education domain. Recent findings in the literature suggest that strategic use of code-switching of bilinguals L1 and L2 in instruction serves multiple pedagogic functions across lexical, cultural and cross-linguistic dimensions, and could enhance students bilingual development and maximize their learning efficacy [9, 10]. This seems to be a particularly effective strategy especially when instructing low proficient language learners [11]. Therefore, the understanding of code-switched speech and development of computational tools for automatically processing such language would provide an important pedagogic aid for teachers and learners in classrooms, and potentially even enhance learning at scale and personalized learning.

Researchers have made significant progress in the automated processing of code-switched *text* in recent years. While Joshi [12] had already proposed a formal computational linguistics framework to analyze and parse code-switched text in the early eighties, it was not until recently that significant strides were made in the large-scale analysis of code-switched text. These have been facilitated by burgeoning multilingual text corpora (thanks largely to the rise of social media) and corpus analysis studies (see for example [13, 14, 15]), which have in turn facilitated advances in automated processing, including part-of-speech tagging [16, 17], predicting code-switch points [18], and language identification [19, 20].

There is comparatively less work in the literature on automated analysis of code-switched *speech*, partially due to the relative lack of structured corpora (as compared to those for text-based work) and also potentially because it also poses yet another significant challenge in the form of speech recognition for multiple languages. Nonetheless, some researchers have made strong strides in spoken corpus development to support such research in certain language pairs, for instance, Mandarin–English [21, 22], Cantonese–English [23] and Hindi–English [24], which have in turn led to developments in automatic speech recognition [25, 26] and language modeling [27]. However, these are limited; there remains a need for more code-switched speech resources in these and other languages to spur research into the automated processing and analysis of such data.

Given that this is still a growing field, it is perhaps understandable why there is limited or no research on the automated analysis of conversational, code-switched *dialog*, let alone the building of bilingual dialog systems that are capable of code-switching. An important step to achieving this involves the collection of a large corpus of code-switched spoken dialog, that can be used for training and analysis. This paper is a first attempt to bridge this gap, to our knowledge. We present a multimodal corpus of human–machine code-switched dialog in both English–Hindi and English–Spanish that can be leveraged for code-switching research. The data collection framework, described in the following section, leverages an open-source spoken dialog system in a crowdsourced paradigm to obtain conversational speech data from a large number of speakers.

2. The HALEF dialog ecosystem

We use the open-source HALEF (Help Assistant – Language-Enabled and Free) dialog system[1] to develop conversational applications within the crowdsourcing framework (see Figure 1). Where there are multiple academic (Olympus [28], Alex [29], Virtual Human Toolkit [30], OpenDial[2], etc.) and industrial (Voxeo[3], Alexa [4], etc.) implementations of spoken and multimodal dialog systems, many of these often use special ar-

[1] http://halef.org.
[2] http://www.opendial-toolkit.net
[3] https://voxeo.com/prophecy/
[4] https://developer.amazon.com/alexa

Table 1: *Dialog system prompts for the conversational code-switched items.*

Turn	English	English–Hindi	English–Spanish
1	Hi, welcome to The Coffee Spot. What can I get you today?	Hi! Coffee Spot me aapka swaagat hai! Would you like something to drink?	Hola! Bienvenido al Coffee Spot. Would you like something to drink?
2	Okay, Is that for here or to go?	Achha, got it. Would you like it for here ya phir parcel lenge?	Muy bien, I got it. Would you like it for here o para llevar?
3	Okay, would you like that hot or iced?	And would you like that thanda ya garam?	And would you like that frio o caliente?
4	And did you want that drink to be small, medium, or large?	OK. Aur aapko small, medium ya large size chahiye?	OK. Que tamaño lo quiere, small, medium or large?
5	And would you like that with milk or sugar/cream?	And would you like that with doodh ya cheeni?	And would you like that con crema o leche?
6	Perfect. Did you want something to eat with that?	Theek hai, perfect. Aur aapko kuch aur chahiye tha, to eat with that?	Muy bien, perfect. Le gustaria algo mas de comer, with that?
7	And I'm assuming you'd like that toasted?	And I'm assuming ki aapko woh toasted chahiye, right?	And I'm assuming le gustaria tostado?
8	Alright, thanks. Your order will be out shortly.	Okay ji. Your order will be ready shortly.	Gracias. Your order will be ready shortly.

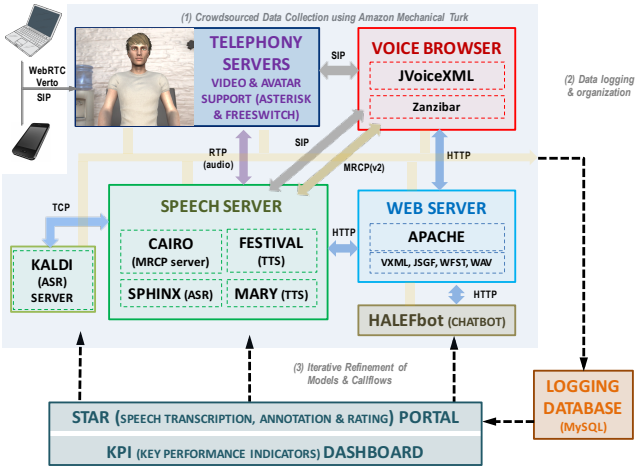

Figure 1: *The HALEF dialog system used in a crowdsourcing-based iterative bootstapping setup for rapid development and data collection.*

chitectures, interfaces, and languages paying little attention to existing speech and multimodal standards (see [31] for more details). In comparison, HALEF is an open-source, modular, cloud-based dialog system that is compatible with multiple World Wide Web Consortium (W3C) and open industry standards. The HALEF architecture and components have been described in detail in prior publications [31, 32].

3. Data

3.1. Conversational item design

We hypothesized that using *intra*sentential code-switched prompts would increase the probability of eliciting code-switched responses (either inter- or intrasentential) from callers. In addition, since code-switching is often contingent upon the socio-pragmatic setting under consideration, we chose a more informal task domain, a coffee shop interaction, as the basis of our conversational task. In this task, we asked callers to pretend that they were in a coffee shop and order a drink and a food item from a provided menu. The automated system plays the role of a barista who takes their order. In order to avoid complications with the language understanding for this initial prototype, we kept the dialog flow non-branching, i.e., the system moved on to the next prompt irrespective of what users said. The ultimate aim of such a task template is to provide speaking practice and, potentially, interactive feedback to language learners. Table 1 lists the dialog flow of the task along with the specific prompts used for both the English–Hindi and English–Spanish cases.

Table 2: *Corpus statistics.*

Item	English–Hindi	English–Spanish
Number of calls collected	555	150
Number of calls transcribed	200	110
Number of unique tokens transcribed	English: 2195 Hindi: 2274	English: 1175 Spanish: 1338
Utterance-level language use or codeswitching percentage	English: 32% Hindi: 35% Both: 33%	English: 36% Spanish: 51% Both: 13%
Gender distribution	82% male	67% male
Self-described daily language use preference	English : 23% Hindi : 21% Either : 56%	English : 34% Spanish : 11% Either : 55%

3.2. Crowdsourcing data collection

We used Amazon Mechanical Turk for our crowdsourcing data collection experiments. Crowdsourcing (particularly via Amazon Mechanical Turk) has been used in the past for the assessment of dialog systems as well as for collection of interactions therewith [33, 34, 35]. We leveraged the aforementioned HALEF dialog system to develop conversational applications within this crowdsourcing framework and collect data from Amazon Mechanical Turk workers. In this iterative data collection framework, depicted schematically in Figure 1, the data logged to the database during initial iterations is transcribed, annotated, rated, and finally used to update and refine the conversational task design and models (for speech recognition, spoken language understanding, and dialog management). In addition to calling into the system to complete the conversational tasks, callers were requested to fill out a 2-3 minute survey regarding different aspects of the interaction, such as their overall call experience, how engaged they felt while interacting with the system, how well the system understood them, to what extent system latency affected the conversation, etc.

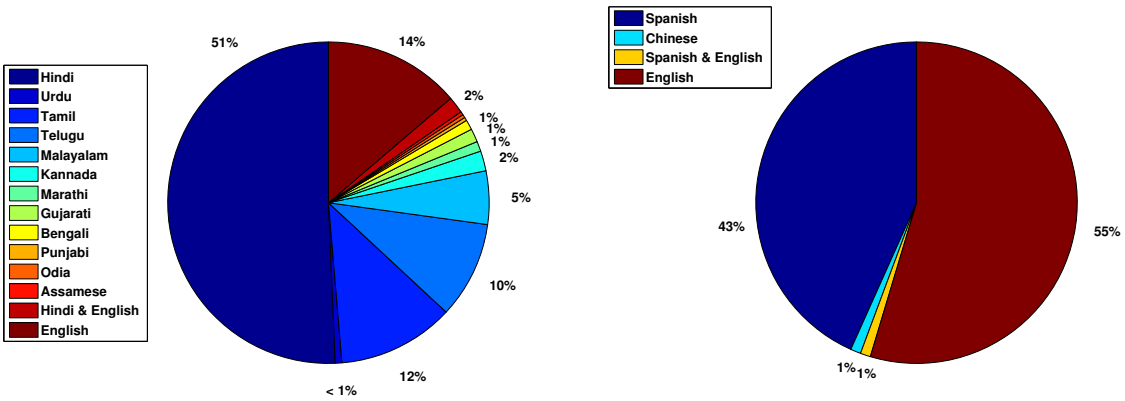

Figure 2: *Percentages of callers of different first languages in the English–Hindi (left) and English–Spanish (right) corpora.*

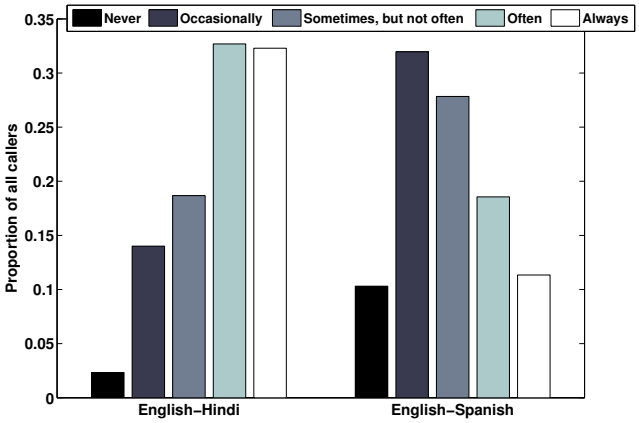

Figure 3: *Callers' self-reported code-switching tendencies.*

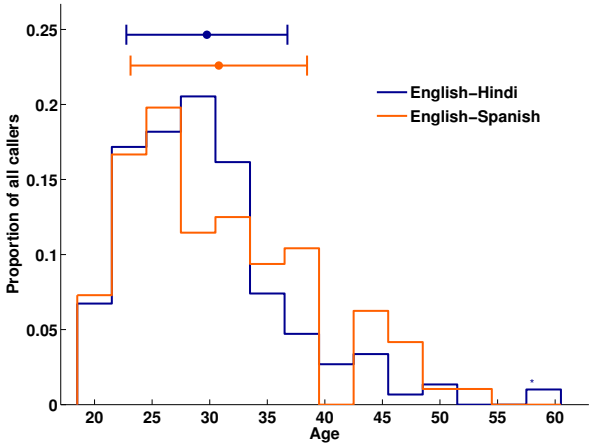

Figure 4: *Age distributions of callers for both dialog items.*

3.3. Corpus Characteristics

Table 2 briefly summarizes the primary statistics of the English–Hindi and English–Spanish corpora collected thus far. While we collected over 700 calls in all, we only transcribed a subset of the corpus (resulting in more than 4400 and 2500 transcribed tokens for the English–Hindi and English–Spanish corpora, respectively) owing to time and resource constraints[5]. Nonetheless, performing this exercise offers us a window into the richness of the data collected; we list some of the insights obtained as a result of this in Section 4.

More than 50% of callers described themselves as having no particular preference for English or Hindi (or Spanish) in daily conversations. This is particularly encouraging, as we would like to collect speech from speakers who are fluent in both languages being considered. Looking at the distributions of first or native languages (L1s) across the speaker population is also particularly revealing – while speakers who called into the English–Spanish item primarily reported Spanish or English as their L1, their English–Hindi counterparts were comparatively more multilingual, with a lot more callers reporting a native tongue other than Hindi and English. This is commonplace in India, where a large swathe of the country speaks English and Hindi, while also speaking the state language. Furthermore, when asked to report how often they used code-switched language in daily life, most English–Hindi callers replied they either did so often, if not always (see Figure 3). A smaller, but still substantial proportion of English–Spanish speakers also reported of a tendency to code-switch during daily conversations, and while this is reflected in the comparatively smaller percentage of code-switched English–Spanish utterances in the corpus as compared to English–Hindi, the nature of our specific crowdsourcing speaker pool could also be an significant contributing factor here. While speakers were primarily male, they distributed across a wide age range (see Figure 4). In addition, the average handling time was around 100 seconds for both the English–Hindi and English–Spanish codeswitched items – see Figure 5 for a histogram distribution of call durations.

[5]Having said that, the iterative HALEF data collection framework allows us to continue collecting many more calls, and transcribing them going forward.

4. Observations and Analysis

4.1. Qualitative Analysis of Caller Responses

People tended to use different strategies in responding to the code-switched machine prompts. We enumerate a few of these below:

1. *Reinforcement or repetition*: Saying the same thing in two languages to ensure that they got the message across. Examples include:
   - `<Hi>Uh mujhe to thanda pasand hai.</Hi><En>I'll go for cold.</En>`
   - `<Sp>Quiero grande.</Sp> <En>Large please</En>.`

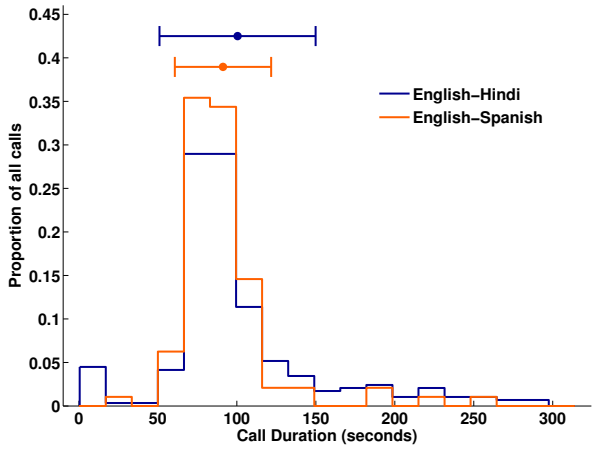

Figure 5: *Histograms of call durations for both dialog items.*

2. *Both inter- and intra- sentential code-switching*: This often happened within the same interaction, over the course of the dialog flow.

    - `<Hi>Uh mujhe</Hi><En>I'll go for large.</Hi>`, followed two turns later by `<En>Um, anything.</En><Hi>Khaane ke aapke paas kya hai?</Hi>`
    - `<En>Eh could you also give me a croissant</En><Sp>por favor</Sp>`

3. *Word-language alternations*: Marked by the use of alternating words in alternating languages. For instance:

    - `<En>With</En><Hi>doodh</Hi> <En>and</En><Hi>cheeni.Dono.</Hi>`

4. *Adjustment strategies*: This was observed among speakers who were multilingual, and neither English or Hindi/Spanish was a first language, where they would code-switch between lexical items in different languages, or use filler words like "yeah" in English followed by a phrase in the other language.

    - `<En>Uh what</En><Hi>Aapke paas kya hai</Hi><En>available?</En>`

5. *Single word responses* to different questions (in both languages) over the course of a single call.

6. *Switching languages gradually* over the course of the call. This could potentially occur when the caller is fluent in both languages, but prefers speaking one over the other.

4.2. Quantitative Analysis of User Experience

Figure 6 shows how users of both the English–Hindi and English–Spanish systems rated different aspects of their call experience on a 1–5 Likert scale (1 being highly unsatisfactory and 5 being extremely satisfactory). This includes how engaged they felt during the interactions and how satisfactory the overall system performance and latency was, along with how well they felt the system understood them (SLU degree). We found the user experience to be overwhelmingly positive, with a large

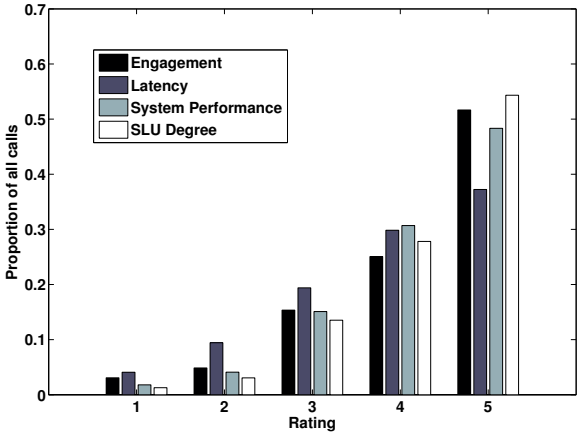

Figure 6: *Distributions of various user experience metrics as rated by all callers in the corpus. See text for details.*

proportion of users assigning ratings of 4 or higher. This is encouraging as we continue to build and develop more engaging code-switched conversational tasks going forward.

5. Discussion and Outlook

We have presented a paradigm for building a corpus of code-switched human–machine dialog using the HALEF open-source cloud-based dialog ecosystem in a crowdsourcing framework, and have further shown that a conversational item with intra-sentential code-switched prompts is effective in eliciting code-switched responses from bi- or multilingual speakers. Having said that, the current item design is not conducive to eliciting code-switched responses longer than two or three sentences on average, given the relatively pointed nature of the question prompts. While this allows us to elicit responses that are structured and directed which makes early linguistic analysis and corpus study easier, future work will look at designing more complex items that elicit longer and more open-ended responses, and incorporate more branching. Such improvements, as well as the potential scaffolding of multiple code-switched dialog items, could be useful in the development of language learning module for new language learners.

Such a corpus also has much potential for the automated processing of code-switched dialog. Along with the previously-studied problems of automatic speech recognition, language identification, and parsing, such a corpus presents a resource for studying other important issues, such as spoken language understanding of code-switched speech, dialog management, and perhaps, with future generations of such work, even code-switched language generation. Finally, besides its value to our future automation endeavors, we also envision the corpus being a useful resource for the linguistic analysis of code-switched dialog observed in everyday conversational settings.

6. Acknowledgements

We thank Keelan Evanini, Eugene Tsuprun and Juan Manuel Bravo for useful discussions regarding the design and development of the codeswitched dialog items and Rutuja Ubale, Robert Pugh and Juan Manuel Bravo for help with transcribing the data. We are also grateful to Saerhim Oh, Larry Davis, Veronika Timpe-Laughlin, Tanner Jackson, Pablo Garcia Gomez, John Norris, and Spiros Papageorgiou, who contributed to the original item based on which the codeswitched item was designed.

7. References

[1] L. Milroy and P. Muysken, *One speaker, two languages: Cross-disciplinary perspectives on code-switching*. Cambridge University Press, 1995.

[2] S. Poplack, "Sometimes ill start a sentence in spanish y termino en español: toward a typology of code-switching1," *Linguistics*, vol. 18, no. 7-8, pp. 581–618, 1980.

[3] D. Sankoff and S. Poplack, "A formal grammar for code-switching," *Research on Language & Social Interaction*, vol. 14, no. 1, pp. 3–45, 1981.

[4] E. Woolford, "Bilingual code-switching and syntactic theory," *Linguistic inquiry*, vol. 14, no. 3, pp. 520–536, 1983.

[5] P. Muysken and L. Milroy, "Code-switching and grammatical theory," *Onse speaker, two languages*, pp. 177–198, 1995.

[6] J. MacSwan, "Code switching and grammatical theory," *The handbook of bilingualism*, vol. 46, p. 283, 2004.

[7] L. Wei, *The bilingualism reader*. Psychology Press, 2000.

[8] C. Myers-Scotton, "Codeswitching with english: types of switching, types of communities," *World Englishes: Critical Concepts in Linguistics*, vol. 4, no. 3, p. 214, 2006.

[9] R. S. Wheeler, "Code-switching," *EDUCATIONAL LEADERSHIP*, 2008.

[10] Y.-L. B. Jiang, G. E. García, and A. I. Willis, "Code-mixing as a bilingual instructional strategy," *Bilingual Research Journal*, vol. 37, no. 3, pp. 311–326, 2014.

[11] B. H. Ahmad and K. Jusoff, "Teachers code-switching in classroom instructions for low english proficient learners," *English Language Teaching*, vol. 2, no. 2, p. 49, 2009.

[12] A. K. Joshi, "Processing of sentences with intra-sentential code-switching," in *Proceedings of the 9th conference on Computational linguistics-Volume 1*. Academia Praha, 1982, pp. 145–150.

[13] T. Solorio, E. Blair, S. Maharjan, S. Bethard, M. Diab, M. Gohneim, A. Hawwari, F. AlGhamdi, J. Hirschberg, A. Chang *et al.*, "Overview for the first shared task on language identification in code-switched data," in *Proceedings of the First Workshop on Computational Approaches to Code Switching*. Citeseer, 2014, pp. 62–72.

[14] K. Bali, Y. Vyas, J. Sharma, and M. Choudhury, "i am borrowing ya mixing? an analysis of english-hindi code mixing in facebook," *Proceedings of the First Workshop on Computational Approaches to Code Switching, EMNLP 2014*, p. 116, 2014.

[15] G. Molina, N. Rey-Villamizar, T. Solorio, F. AlGhamdi, M. Ghoneim, A. Hawwari, and M. Diab, "Overview for the second shared task on language identification in code-switched data," *EMNLP 2016*, p. 40, 2016.

[16] T. Solorio and Y. Liu, "Part-of-speech tagging for english-spanish code-switched text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 1051–1060.

[17] A. Jamatia, B. Gambäck, and A. Das, "Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages." in *RANLP*, 2015, pp. 239–248.

[18] T. Solorio and Y. Liu, "Learning to predict code-switching points," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, pp. 973–981.

[19] U. Barman, A. Das, J. Wagner, and J. Foster, "Code mixing: A challenge for language identification in the language of social media," *EMNLP 2014*, vol. 13, 2014.

[20] B. King and S. P. Abney, "Labeling the languages of words in mixed-language documents using weakly supervised methods." in *HLT-NAACL*, 2013, pp. 1110–1119.

[21] Y. Li, Y. Yu, and P. Fung, "A mandarin-english code-switching corpus." in *LREC*, 2012, pp. 2515–2519.

[22] D.-C. Lyu, T.-P. Tan, E.-S. Chng, and H. Li, "Mandarin–english code-switching speech corpus in south-east asia: Seame," *Language Resources and Evaluation*, vol. 49, no. 3, pp. 581–600, 2015.

[23] J. Y. Chan, P. Ching, and T. Lee, "Development of a cantonese-english code-mixing speech corpus." in *INTERSPEECH*, 2005, pp. 1533–1536.

[24] A. Dey and P. Fung, "A hindi-english code-switching corpus." in *LREC*, 2014, pp. 2410–2413.

[25] N. T. Vu, D.-C. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E.-S. Chng, T. Schultz, and H. Li, "A first speech recognition system for mandarin-english code-switch conversational speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4889–4892.

[26] C.-F. Yeh, L.-C. Sun, C.-Y. Huang, and L.-S. Lee, "Bilingual acoustic modeling with state mapping and three-stage adaptation for transcribing unbalanced code-mixed lectures," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5020–5023.

[27] H. Adel, N. T. Vu, and T. Schultz, "Combination of recurrent neural networks and factored language models for code-switching language modeling." in *ACL (2)*, 2013, pp. 206–211.

[28] D. Bohus, A. Raux, T. Harris, M. Eskenazi, and A. Rudnicky, "Olympus: An Open-Source Framework for Conversational Spoken Language Interface Research," in *Proc. of the HLT-NAACL*, Rochester, USA, 2007.

[29] F. Jurčíček, O. Dušek, O. Plátek, and L. Žilka, "Alex: A statistical dialogue systems framework," in *Text, Speech and Dialogue*. Springer, 2014, pp. 587–594.

[30] A. Hartholt, D. Traum, S. C. Marsella, A. Shapiro, G. Stratou, A. Leuski, L.-P. Morency, and J. Gratch, "All together now," in *Intelligent Virtual Agents*. Springer, 2013, pp. 368–381.

[31] V. Ramanarayanan, D. Suendermann-Oeft, P. Lange, R. Mundkowsky, A. V. Ivanov, Z. Yu, Y. Qian, and K. Evanini, "Assembling the Jigsaw: How Multiple Open Standards Are Synergistically Combined in the HALEF Multimodal Dialog System," in *Multimodal Interaction with W3C Standards*. Springer, 2017, pp. 295–310.

[32] Z. Yu, V. Ramanarayanan, R. Mundkowsky, P. Lange, A. Ivanov, A. W. Black, and D. Suendermann-Oeft, "Multimodal halef: An open-source modular web-based multimodal dialog framework," in *International Workshop on Spoken Dialog Systems (IWSDS 2016), Saariselka, Finland*, 2016.

[33] I. McGraw, C.-y. Lee, I. L. Hetherington, S. Seneff, and J. Glass, "Collecting voices from the cloud." in *LREC*, 2010.

[34] F. Jurcıcek, S. Keizer, M. Gašic, F. Mairesse, B. Thomson, K. Yu, and S. Young, "Real user evaluation of spoken dialogue systems using amazon mechanical turk," in *Proceedings of INTERSPEECH*, vol. 11, 2011.

[35] V. Ramanarayanan, D. Suendermann-Oeft, P. Lange, A. V. Ivanov, K. Evanini, Z. Yu, E. Tsuprun, and Y. Qian, "Bootstrapping development of a cloud-based spoken dialog system in the educational domain from scratch using crowdsourced data," *ETS Research Report Series*, 2016.