Comparing Command Construction in Native and Non-Native Speaker IPA Interaction through Conversation Analysis

Yunhan Wu
University College Dublin
Ireland
yunhan.wu@ucdconnect.ie

Martin Porcheron
Swansea University
Wales, UK
m.a.w.porcheron@swansea.ac.uk

Philip R. Doyle
University College Dublin
Ireland
philip.doyle1@ucdconnect.ie

Justin Edwards
University College Dublin
Ireland
justin.edwards@ucdconnect.ie

Daniel Rough
University of Dundee
Scotland, UK
drough001@dundee.ac.uk

Orla Cooney
University College Dublin
Ireland
orla.cooney@ucdconnect.ie

Anna Bleakley
University College Dublin
Ireland
Anna.Bleakley@ucdconnect.ie

Leigh Clark
Swansea University
Wales, UK
l.m.h.clark@swansea.ac.uk

Benjamin R. Cowan
University College Dublin
Ireland
benjamin.cowan@ucd.ie

Abstract

Intelligent Personal Assistants (IPAs) are limited in the languages they support, meaning many people are left to interact using a non-native language. Yet, we know little about how people interact with IPAs in this way. Through a conversation analysis (CA) perspective, we examine native (L1) and non-native (L2) English speaker interactions with Google Assistant, comparing how both user groups produce IPA commands. Our work shows that L1 and L2 speakers similarly used pauses, partial or complete repetition, and hyper-articulation when constructing commands. However, L2 speakers tended to experience issues in lexical access, syntactic construction and pronunciation, resulting in the use of code-mixing, increased pause lengths and off-task rehearsal to help generate commands. We consider reasons for such effects, whilst exploring ways to design IPA interaction to ensure it is sensitive to L2 challenges in command production.

CCS Concepts

• **Human-centered computing** → **User studies**; **Natural language interfaces**; *Accessibility design and evaluation methods.*

Keywords

speech interface, voice user interface, intelligent personal assistants, non-native speakers

ACM Reference Format:
Yunhan Wu, Martin Porcheron, Philip R. Doyle, Justin Edwards, Daniel Rough, Orla Cooney, Anna Bleakley, Leigh Clark, and Benjamin R. Cowan. 2022. Comparing Command Construction in Native and Non-Native Speaker IPA Interaction through Conversation Analysis. In *4th Conference on Conversational User Interfaces (CUI 2022), July 26–28, 2022, Glasgow, United Kingdom.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3543829.3543839

CUI 2022, July 26–28, 2022, Glasgow, United Kingdom

1 Introduction

Due to the growth in popularity of Intelligent Personal Assistants (IPAs) such as Google Assistant and Amazon's Alexa, using speech as a means of interface interaction is now well established [12]. These IPAs support a number of languages, yet they are by no means comprehensive in the functionality offered across all supported languages (e.g., [37]). Consequently, some users are forced to either interact in a non-native language (i.e., they must interact with IPAs as L2 speakers), or be excluded from some IPA functionality entirely.

Recent efforts have been made to explore non-native speaker IPA user experience [53, 62, 63], yet we currently know little about how L2 speakers behave linguistically, and how this varies from users who use their native language (L1 speakers) when interacting with IPAs. In this study, we aim to bridge this gap by identifying key similarities and differences in the speech and language patterns of L1 and L2 speakers respectively. To achieve this, we designed a study where L1 and L2 English speakers completed a set of scenarios with Google Assistant on a smartphone and smart speaker, after which they took part in a semi-structured interview to reflect on their interaction session. Conversation analysis (CA) of the interaction data, combined with semi-structured interview data used to guide our interpretations, found that L1 and L2 speaker interactions shared similar linguistic attributes such as pausing mid-command when unsure of how to formulate utterances, using *second saying* (i.e., partial or full repetition of an utterance within the same conversational turn [60]) or repetition of a command when the IPA was not responsive in a successive turn. Like L1 speakers, L2 speakers also used hyper-articulation in an attempt to improve command recognition. Unique to speech by L2 speakers were challenges in producing the complete command before the device stopped 'listening', difficulties in identifying the correct words

and structures for commands, and perceived issues with pronunciation. L2 speakers also tended to 'code mix', whereby they would combine words and utterances from their native language in their L2 speech. L2 speakers also used unique strategies in approaching interaction such as breaking tasks into simpler 'sub-commands' and off-task rehearsal. Importantly, these effects were observed during use of IPAs on both smart speakers and smartphones. Our findings add to the growing interest in L2 speaker interaction by highlighting specific behaviours that occur when people interact with IPAs using a second language, emphasising linguistic differences in command construction across L1 and L2 speakers that need to be considered when designing IPAs for L2 use. Based on our findings, we suggest that IPA design should be made more sensitive to issues in L2 language production, as well as supporting the command generation process.

2 Related Work

2.1 Intelligent Personal Assistant Interaction

Although there have been claims that machines have reached parity with humans in conversational speech recognition [64], IPAs such as Google Assistant and Amazon Alexa are still mostly used for simple user-led tasks. Their primary functions are to conduct information searches, interact with Internet of Things (IoT) devices, play music, and set alarms and timers [4, 42]. These tasks tends to be executed using limited question-answer type dialogues with a limited number of turns [31, 51]. Recent work has identified a number of key challenges when interacting with IPAs revolving around user trust in executing more complex or socially sensitive tasks. For example, sending a message or calling a contact [43], accurately recognising users' accented speech [20], as well as issues in the use of human-like design choices (e.g., human-like voices and linguistic content) inaccurately portraying an IPA's actual capability [13, 20, 27, 43]. More recently, users have emphasised privacy, as well as data collection and access practices [4, 14, 20], as major challenges to be addressed within IPA-based interaction.

When interacting with IPAs, users tend to use particular linguistic patterns including altering lexical choice (e.g., using simpler terms [43, 51]), shortening or using command and keyword-like utterances [51] as well as enunciating more clearly, hyperarticulating, and altering their accent when interaction breakdowns occur [43, 51]. Similar effects are seen when comparing human-machine dialogue (HMD) to human-human dialogue (HHD). When interacting with machines, people regularly use fewer anaphora and coherence markers [2], shorter utterances, simpler words, and less complex grammatical utterances [8, 36] and tend to hyperarticulate when encountering errors [48], all in the aim of ensuring the can 'recognise' their input [55]. This adaptation of speech has been attributed to people's perceptions of a machine's abilities [11, 39], termed as our 'partner models' [26]). These represent people's perceptions of a machine's communicative competence and flexibility, and how human-like systems are in the way they communicate. Generally, people tend to see machine dialogue partners as having limited capabilities [9], being seen as "at risk listeners" [48]–that is, interlocutors at high risk of communicative failure. We thus adapt our utterances to more likely ensure communicative success [10, 19] and revise our utterances in cases of failure [30]. This adaption is not, however, exclusively based on the perceived limitations of systems. Research also shows people consider their own speech and language behaviours (i.e., accent, pronunciation and speech rate) when engaging in behavioural affordances aimed at improving the chances of successful communication with speech interfaces [27, 43, 51].

2.2 Non-native speakers' interaction with IPAs

Due to the lack of language coverage and the variability of functionality across IPAs and their supported languages (e.g., [37]), many users speak to IPAs in a language that is not their 'mother tongue'. As such, recent work has focused on the challenges faced by users who have to use languages other than their first language to interact with IPAs. When comparing native (L1) and non-native English speakers' (L2) user experience of IPAs on mobiles and smart speakers, Wu et al. found that L2 users appreciated visual feedback afforded by mobile-based IPA use [63]. Visual feedback allowed L2 users to diagnose errors and reassured them that the system understood their commands accurately, whereas L1 speakers felt less of a need for such visual feedback to support their interaction [63]. This may go some way to explaining why L2 speakers find smart speakers harder to use [53, 54] and more difficult to interact with effectively, than L1 speakers [53, 54].

A major difficulty perceived by L2 speakers lies specifically in language generation and production during IPA interaction [63]. L2 speakers tend to need more time to plan [54] and produce speech, as well as interpret the system's utterances [63]–requirements that current IPAs are not sensitive to [63]. This echoes research that suggests L2 speakers can face challenges in language production in situations where they lack the knowledge of a non-native language [25, 59]. L2 language speakers also commonly feel like they experience issues with lexical retrieval–a process that is less automatic and more effortful when generating utterances in a non-native tongue [21, 32, 57]. While L1 speakers focus on generating concise commands during interaction, L2 speakers tend to perceive that they pay more attention to their pronunciation during IPA use [63]. The need to rephrase commands when errors occur is also a cause of frustration for L2 speakers [54], potentially due to the lexical retrieval and language generation issues identified previously. The perceived linguistic difficulties L2 speakers face when interacting with IPAs are consistent with findings that L2 speakers experience higher mental workload than L1 speakers when interacting with IPAs [62]. Although these linguistic difficulties are highlighted in previous work, observations have been based on reflections in interviews or through purely quantitative assessment of perceptions and commands that users generate when interacting with IPAs. Our work here contributes by taking a Conversation Analysis approach to qualitatively observe whether and how these language generation patterns actually occur within IPA interaction.

2.3 Analysing IPA interaction using Conversation Analysis

Conversation Analysis (CA) [56] has a long history in the analysis of user language to guide the design of speech interactions with computers (e.g., [61]). The perspective enables researchers

to identify and analyse the numerous different approaches people adopt in HMD interaction, based specifically on their speech and language behaviour. Yet, rather than focusing on the frequency of language-based phenomena, CA allows for a richer and more in-depth exploration of the effects seen in utterances during the interaction, focusing on significant fragments of interest to illustrate trends within the data. CA has been used to document the ways that IPAs on portable devices become occasioned and used within conversations amongst friends in public spaces like cafés [52], how use of smart speakers becomes an embedded activity within the home [51], and how people converse through voice with a robot interface [49]. Through this work, CA is continually shown to be a robust analytic and generative approach for exploring user interaction with IPAs within various settings. Concepts from CA have also been imported into analyses and discussions about the use of IPAs, including "conversational UX design" [45]. Specific examples include notions of *recipient design* [5] (i.e., the ways in which people formulate their utterances for their recipients) and *progressivity* [30], the idea that people will work to resolve or progress a conversation. Others have used CA to critique existing conversational design features including wake words [1], the tensions between IPA and conversational design, and existing conceptualisations of 'conversation' [50].

3 Research aims and contribution

Recent work on L2 language production has typically taken a quantitative approach (e.g., [62]), which can be reductive and mask patterns of speech that can be more clearly seen through in-depth qualitative observation of dialogues. Other recent work has relied on perceptions that L2 speakers have about language production challenges in IPA interaction [63], without identifying whether there is evidence of these in command construction. Our study here contributes by adopting a CA approach for an in-depth exploration that compares language patterns among L1 and L2 speakers in user-IPA dialogues. From this, we contribute much-needed knowledge of how L1 and L2 speech converges and diverges when interacting with IPAs, based on observations of actual interactions. By observing these patterns, we aim to inform future IPA design towards the creation of more inclusive and effective interactions for L2 users.

4 Study Method

Data used in this research was gathered as part of a larger study on L1 and L2 language speakers' experiences of IPA interaction [62, 63]. Work previously published from this data emphasised quantitative findings, and focused on a narrow set of specific variables. Due to limitations of quantitative approaches outlined above, here we re-examine the data through a broader, richer lens offered by CA. The method of this study, relevant to the data included in this work, is included below.

4.1 Participants

33 participants were recruited for this study (F=14, M=18, Prefer not to say=1; Mean age=28.1 years, SD=9.8). Participants were recruited from a European university via email, posters displayed across the campus, and snowball sampling and were given a €10 voucher as an honorarium. While 33 participants were initially recruited, one participant was removed from the sample due to a technical failure in the study recording process. This left 32 participants for the final analysis. Of these 32, 16 were native English speakers (F=8, M=7, Prefer not to say=1) and 16 were native Mandarin speakers using English as their second language (F=6, M=10).

78.1% (N=25) of all participants mentioned they had experience of using IPAs, including 13 native Mandarin speakers and 12 native English speakers. 3 participants indicated frequent use of IPAs. Apple's Siri (56%) was the most popular IPA used by participants, followed by Amazon Alexa (36%) and Google Assistant (12%). Using a 7-point Likert scale question (1 = Not at all proficient; 7 = Extremely proficient), the 16 Mandarin speakers rated themselves as having medium levels of English proficiency (M=4.21, SD=0.7).

4.2 Device conditions

For the study, participants interacted with Google Assistant through both a Moto G6 smartphone (*Smartphone* condition) and a screenless Google Home Mini smart speaker (*Smart speaker* condition) in a within-participants design. The order in which participants interacted with devices was counterbalanced across participant groups. In both device conditions, participants were asked to only use speech as a way of interacting with the devices. To ensure that the interactions realistically reflected those that occur with each device, feedback mechanisms of each device were preserved (i.e., smartphone screen and voice-based feedback for Google Assistant; visual status lights and speech feedback for Google Home). We note that, although we include different device conditions within the study, the work presented identifies, and thus focuses on, similar linguistic patterns across both device conditions. This allows us to emphasise the commonalities in linguistic patterns when interacting across these two common device types.

4.3 Scenarios

Participants were asked to interact with the two Google Assistant devices in the completion of 12 scenarios (six scenarios per device). These scenarios were based on common activities conducted with IPAs [4, 28] and involved participants (1) playing music, (2) setting an alarm, (3) converting values, (4) asking for the time in a particular location, (5) controlling device volume and (6) requesting weather information. Two versions of each scenario were generated to create two sets of six scenarios. Each scenario was presented to participants as a pictogram (see Figure 1 and supplementary material). This was done to eliminate the potential influence of written instructions on what participants may choose to say to the IPAs, and thus more closely represent natural query generation. In turn, the hope was to reduce potential inconsistencies in translating written queries for L2 speakers whilst also allowing the researchers to have a clear idea of the task that the user was attempting to perform. Scenario sets were counterbalanced across both device and group conditions, with scenario order also randomised within sets for each participant.

4.4 Interview

Following the study, participants took part in a semi-structured interview. The interview focused on (1) general attitudes towards

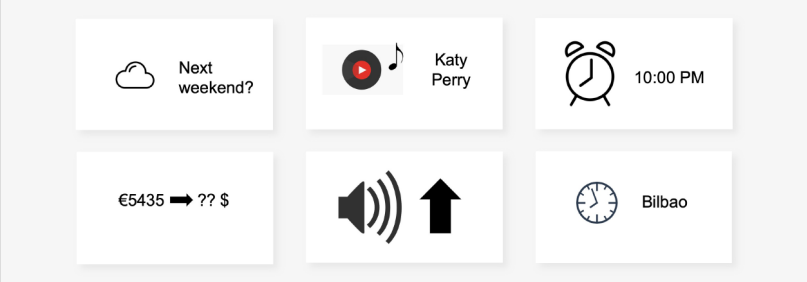

Figure 1: Example set of scenario pictograms

IPAs, (2) their experiences with the IPAs during the study, (3) reflections on how they spoke to the IPAs in the study. Interviews lasted approximately 20 minutes. To prevent linguistic barriers, L2 English speakers had their interviews conducted in Mandarin. Audio data were recorded and transcribed, with L2 English speaker interviews being translated from Mandarin to English by a Mandarin-speaking member of the research team. The interview data is used to help us to more effectively interpret the effects seen, using relevant quotes to inform our interpretations of the fragments in Section 5.

4.5 Procedure

The research received ethical approval through the University's low-risk project ethics procedure. Before starting the study, participants were given an information sheet detailing the nature of the study and what participation entailed. Once participants indicated they understood the study procedure thoroughly, they were asked to provide written consent. Participants then completed a demographic questionnaire, providing their age, sex, nationality, native language, details about their prior experience using IPAs, and an indication of their English proficiency. Following this, a researcher explained the general task participants had to conduct in the study. Before interacting with the IPAs, participants were given a trial set of scenario pictograms on paper and were asked to write down what they would say to the IPA to complete each scenario depicted. The sample pictograms were similar in topic and layout to the study scenarios but varied in information requested (e.g., includes the name of a different artist's music). The researcher then checked these to ensure that participants had interpreted the pictograms correctly before being given the pictograms to be used in the experimental session. After completing the practice phase, participants were asked to interact with Google Assistant using either a smartphone or smart speaker, completing a set of six pictogram scenarios, shown on a laptop screen one at a time. Participants would self-report the status of the scenario after they felt it had been completed or when they experienced difficulty and decided to skip a particular scenario by using a check box. Scenarios were deemed complete by participants rather than experimenters in order to avoid influencing the interaction. After finishing the set of six scenarios, participants moved on to the next device (either smartphone or smart speaker depending on the initial device used) and completed another set of six scenarios, one at a time. As noted in Section 4.2, when interacting with the devices, they were able to engage with all forms of feedback offered by the device being used. In both device conditions, participants were required to use wake words to commence interactions with Google Assistant. Upon completing both sets of scenarios on both devices, participants took part in a brief semi-structured interview, which focused on their experience interacting with IPAs prior to and during the experiment. The researcher then debriefed participants, informing them of the research aims of the study, where and how the data they provided would be used, and contact information should they have any further queries. Participation took approximately 40 minutes and upon completion of the study participants were then given a €10 voucher as an honorarium/reimbursement for their time.

4.6 Interaction Transcription & Fragment Selection

Interactions were audio recorded, with the audio being transcribed by the authors. Rather than using the automatic transcription produced by Google Assistant, which focuses only on what the device interprets the user as saying, we manually transcribed the user's commands being produced as well as the responses from the IPA to the user. This gave us an accurate and rich representation of what occurred from the user side when generating commands. The transcriptions used notation derived from Jefferson notation [6, 35], noting all participant commands, system responses, disfluencies, hesitations, pauses, and other talk included in the recording (e.g., off-task talk, comments to the researcher). Potential fragments (i.e., examples of interactions within the transcripts) that illustrated command construction effects within the interaction data across smartphone and smart speaker IPA interactions were initially selected by the lead author. These were then discussed with two of the authors who had extensive experience and expertise in conducting conversation analysis and speech interface interaction research. Based on the outcomes of the discussions, a set of agreed fragments were chosen from the transcripts. These fragments are presented in Section 5 below.

5 Findings

Our findings begin with examples of common speech and language behaviours employed by both L1 and L2 speakers when uttering commands to IPAs. We then move on to speech and language behaviours for which L2 speakers diverged from L1 speakers, focusing on the methods employed by L2 speakers as they used the IPAs

to complete the scenarios. Throughout this section we present examples extracted from our corpus, with these fragments of data acting as "vivid exhibits" [7, 22] of the actions highlighted, shedding light on our L2 speakers' efforts to complete the pre-set scenarios. We complement these fragments with excerpts from our post-study interview to support our observations.

As noted in Section 4.2, our work identifies common effects seen across both types of device when interacting with IPAs through speech. The fragments presented denote which device was used in the transcriptions, as 'verbal' responses from either 'GA' (representing the Google Assistant on a smartphone) or 'GH' (representing the Google Home smart speaker). The fragments selected are those that cogently highlight the effect being discussed, although for all effects noted, these occurred across both devices and we have merely selected these fragments.

5.1 Common Approaches Between L1 and L2 Speakers

We saw both L1 and L2 speakers produce commands in much the same way as expected from existing literature (e.g., [51]) using pauses when they are unsure of how to continue command production, partial or complete command repetition, and hyper-articulation when faced with perceived IPA 'misunderstanding' issues.

5.1.1 Pausing While Word Learning Our observations show that both L1 and L2 participants inserted *pauses* into their utterances, especially when demonstrating some uncertainty about how to formulate their command. For instance, pauses were common when L1 speakers attempted to utter an unfamiliar word, suggesting the pauses may be related to difficulties with language generation. To contextualise this, we present our first fragment from an L1 speaker, where a participant was tasked with asking a Google Home what the time was in the city of *Bilbao*.

Fragment 1

```
P01-L1 °i assuming that is a city (.) which i cannot pronounce
        (hh)°
       ok google what time is it in (..) bil bao?
GH     the time in bilbao (.) spain is one thirty five pm
```

In this fragment, the participant (P01-L1, i.e., participant 01, who is an L1 speaker), in a quieter volume (denoted by the °), makes a comment that they are unsure of the pronunciation of the name of the city. The participant also chuckles quietly under their breath prior to commencing their request ((hh)). As they perform their instruction, there is a short pause ((. .)) prior to their utterance of the name *Bilbao*, with a small pause between the first and second syllable. They also employ a rising intonation (?) in the second syllable as they utter the word. Through the prior account of their uncertainty of the pronunciation and incomplete utterance thus far, this mid-sentence pause is demonstrably a sign of *hesitation* and of interactional trouble–this is a moment where the participant is 'word learning' bilbao [34]. Nevertheless, the participant ostensibly performs a request according to their best guess. Crucially, the command was successful, with the Google Home (GH) responding by providing the time in Bilbao.

Post-interaction interviews helped to reveal participants' own understandings of why they relied on mid-command pauses. For L2 speakers in particular, challenges in command generation led to a need for more time to be able to generate their utterances.

> *"I might pause in sentences since I have no idea about how to express. I need some time to think"* [P04-L2]

Compared to L1 speakers, L2 participants specifically highlighted that they tended to pause during command construction to support language production, although this phenomena did occur across both cohorts.

5.1.2 Second Saying in Command Production Both participant cohorts also exhibited a tendency to repeat part of their utterances. In the next fragment, in this case from an L2 speaker, the participant asks Google Home for the time in *Helsinki*. As they do so, they pause mid-instruction before repeating some of the words in their initial command.

Fragment 2

```
P07-L2 hey google (.) what's the time (.) in (.) what's the time
       now in helsinki
GH     the time in helsinki (.) finland is four fifty-nine pm
```

Again, the instruction is ostensibly successful, with Google Home providing the time in Helsinki, despite the participant's pauses, restarting, and partial repetition of their command. Our corpus is replete with this form of pause and restart from both L1 and L2 speakers.

This *second saying* [60] is also common in human-human dialogue [60]. In our example above and in other cases in our data set, the IPA demonstrates an ability to respond in the manner desired, despite occurrences of second saying by users. However, instances of second saying did not always lead to successful communication. Participants raised this point in the post-study interviews as a challenge they face that they typically can overcome in human-human dialogue but may struggle with in IPAs, resulting in them having to repeat an entire command.

> *"...and when you speak to a person, you can question closely. For example, if I say something wrong, I can add something after I finish a sentence, even just a single word. But when I'm interacting with IPAs, I need to re-organize words to ask them again."* [P01-L2]

Indeed, as shown in the following section, although modern IPAs can cope with some disfluencies some of the time, their ability to do so is not always consistent. This can result in participants having to repeat entire commands on occasions where the IPA does not respond as desired.

5.1.3 Complete Command Repetition, Hyper-articulation & Uncertainty Participants tended to repeat their entire command in cases where an IPA seemingly did not respond, or did not respond as expected. In the next fragment, the participant issues an instruction to the Google Assistant via a smart speaker to turn the volume down. As there was no audible response from the device, the participant repeated their entire command again.

Fragment 3

```
P02-L1 okay google (.) turn the volume down
       (3.0)
       TURN (.) THE VOLUME (.) DOWN
GH     ((beeps))
```

Lexically speaking, P02-L1's initial command is perfectly formed. However, due to the lack of a response from the device, they appear to treat the request as failed, and repeat the entire command. Through this repeated command, the participant is responding to the device's seeming inability to recognise their prior attempt, and so they repeat the instruction using the same lexical construction. Crucially, this repetition is undertaken at a louder volume, with short pauses between words, with the result of this second instruction treated as a success. This could be demarcated as a form of *hyper-articulation* in which the L1 speaker repeats their request, suggesting that the participant identified the issue as one of the device not 'hearing' the request rather than an issue with their own lexical formulation. This effect is similarly seen in L2 interaction. Consider the singular utterance below from an L2 speaker, in which they are issuing an instruction for the same scenario as in 5.1.1 above.

Fragment 4

```
P02-L2 what's the time now in:: (0.7) bil↑bao (0.4) bilba?
```

In this case, the participant both repeats the place name (bilbao) with a pause between utterances, and has a rising intonation as they do so. This rising intonation in combination with the repeated word seems to suggest that the speaker is unsure as to whether they have pronounced the city correctly, hence the rising intonation which frames the word Bilbao as a question.

From post-interaction interviews, it is clear that both L1 and L2 speakers focus on generating clear speech, being careful to articulate their commands clearly and precisely:

"well I was trying to articulate clearly and you know think about how was, what was the simplest way possible to ask the question you know" [P04-L1]

"So I try to speak formally, pronounce accurately, then, speak slower than normal." [P13-L2]

Reflecting on other work [43, 48], the purpose of this hyper-articulation was apparent to both sets of participants. L1 and L2 participants emphasised how, when they encountered errors, they would hyper-articulate and aim to speak more slowly and loudly so as to be understood by the system:

"I spoke kind of slowly and after it didn't understand me the first time I started trying to speak more clearly and loud." [P06-L1]

"You cannot speak fast. When you speak slower, you can make the machine react. If you speak fast it cannot give you some response to your questions." [P03-L2]

Within L2 speaker interactions, however, direct repetition can persist if the IPA is having difficulty in understanding the command. The following fragment shows the participant issuing a command to complete the scenario of finding out the weather next weekend:

Fragment 5

```
P03-L2 hey google (.) ((murmur)) what weather in the next weekend?
       °what° (2.0) what weather in the next weekend?
GA     here's some result for the search.
P03-L2 what weather in the next wee::kend?
GA     what do we talk about, i lost track
P03-L2 what's the weather in the next wee::kend?
GA     here's some results
```

In this case, the L2 participant repeatedly used a command that was grammatically incorrect. They did not get a response after their first utterance of the command, yet they used the same command, taking a longer pause so as to consider their utterance. The quiet uttering of 'what' suggests that they may be rehearsing pronunciation before restarting their command. What is most notable is the lack of variability or grammatical correction in the command, even after the IPA fails to give the participant the requested information.

While L2 speakers may adopt superficially similar methods (in this case, complete command repetition) a closer look suggests that there may be an increased lack of flexibility due to difficulty in generating other alternative command constructions.

5.2 L2-Specific Approaches

When reviewing the L1 and L2 speaker data above, we see similar phenomena in command construction such as the use of pauses and command repetition. We also observe issues when constructing commands that seem unique to L2 speakers, most notably with time needed for language production and *code mixing*, as well as exacerbated pronunciation challenges.

5.2.1 Producing Language in Time Language production is perceived as a significant issue by speakers using their non-native language in IPA interaction [63]. Across both smartphones and smart speakers, this severely impacted the success of the dialogues. Our next fragment below exemplifies the types of difficulties that many L2 participants had when attempting to construct commands with more complex lexical or technical terms. In this task, the participant is shown a pictogram asking them to convert the temperature '32°C' into '°F'.

Fragment 6

```
P11-L2 hey google (.) how's (0.3) thirty two:: celsius degree (.)
       count in other way
GA     check out these results
P11-L2 hey google (.) how thirty two:: celsi degree counts in other
       way
GA     here's what i found on the web
```

In this fragment we see that participant P11-L2 struggles to identify the name of the unit *Fahrenheit*, which was needed for the temperature conversion command. Using the phrase "in other way" was a secondary strategy to make the IPA realise that they are looking to convert the unit of measurement mentioned in the command. There is also a clear pause between "Celsius degree" and "count in the other way", which may signal the participant's attempt to retrieve appropriate words and generate an alternative phrasing. The fragment also illustrates the struggles that L2 speakers sometimes have with constructing syntactically accurate commands using an unconventional construction for the command, such as "how's thirty two Celsius degree count in other way". The fragment also shows the user potentially diagnosing the issue as one of pronunciation and subsequently changing how they pronounced Celsius between their first and second command utterances.

Although the above fragment was on a smartphone, we also noticed similar issues when L2 speakers interacted with Google

Home, with the participant struggling to find the words for Celsius and Fahrenheit, replacing these with "the temperature" and "another form" respectively in the command.

Fragment 7

```
P05-L2 okay google please help me to:: (0.2) eh exchange the form
       of the :: (.) temperature (0.3) from:: thirty two to::
       (0.2) another form
GH     sorry (.) i'm not sure how to help with that but i'm still
       learning
P05-L2 eh (2.1) chan- (.) change the form of the (1.0) °sorry°
       (1.1) hey (0.4) eh::-
GH     on the website dictionary.com (.) they say:to change
       inform (.) appearance (.) or structure (.) metamorphose
       (.) to change in condition (.) nature (.) or character
       (.) convert-
```

After the Google Home responds without the conversion, the participant seems to display some confusion about how to proceed, using filled pauses, hesitations (eh) and then issuing an apology to the researcher (°sorry°). This follows a long pause where they seem to be looking to generate an alternative command to complete the task (change the form of the) but seem unable to confidently select the correct words. Before they have the chance to generate and produce a second utterance, the Google Home responds with an utterance that is irrelevant to the desired task, suggesting an unintended activation of the device occurred.

These fragments not only emphasise the difficulties that L2 speakers have when generating the needed lexical content and accurate syntactic structure for commands, but also the importance of increased command *generation time* needed for L2 speakers when interacting with IPAs. The following fragment, which sees a participant attempt to complete the scenario of converting a euro amount into dollars, starkly emphasises the lack of time L2 participants have to generate an utterance before the IPA produces a response. The participant's difficulty is only exacerbated by the continual interruptions.

Fragment 8

```
P14-L2 hey google
GA     hey (.) how can I help?
P14-L2 how much does (0.2) five (0.5) thou (0.9) five thousand
       four hundred-
GA          here are some results from the web
P14-L2 hey google (.) how much is five hundred (1.8) hey google
GA                                                  here are
       results from the web
P14-L2 hey google (.) how much is (0.4) five thousand four hundred
       (1.0)
         three (.) thirty five-
GA       here's what i found on the web
P14-L2 hey google (.) how much is five thousand four hundred and
       thirty five euro
GA     five thousand four hundred and thirty five euros equals to
       four thousand five hundred and ninety nine pounds and
       forty two pennies.
P14-L2 hey google (.) hey google (.) how much is four (0.3) five
       hun-
GA     i found these results
P14-L2 hey google (.) how much is (0.2) five thousand (0.5) four
       hundred thirty five euro (1.0)
       equal to::         -
GA           five thousand four hundred and thirty five euros
       equals to four thousand five hundred and ninety nine pounds
       and forty two pennies
P14-L2 hey google (.) how much:: (0.3) dollar (0.3) is equal to
       five thousand four hundred (0.4) and thirty five euro
GA     five thousand four hundred and thirty five euros equals to
       six thousand forty one dollars and eighty seven cents
```

Throughout this fragment, the participant is taking time to produce the correct terms and structure for the command, pausing regularly to facilitate production. Yet the IPA consistently interprets the participant's momentary pauses as the end of their command and consequently responds with the wrong information. Recent work has emphasised that, compared to L1 speech, silent and filled pauses are more frequent and longer within utterances for L2 speech, as these speakers use this time to plan, retrieve and produce language in conversation [24]. The fragment clearly shows a need for voice activity detection used within IPAs to be more sensitive to these pauses and the production time needed for L2 speakers.

Post-interaction interviews further emphasised the challenges that L2 participants face in terms of language generation, particularly in terms of lexical retrieval and grammatical construction:

> *"....when I asked the devices to change the volume, I didn't know the word 'volume' so I changed it to 'voice' or 'sound'."* [P09-L2]

> *"Maybe there are some pronunciation issues....or....some grammar issues. When I say something that rolls right off the tongue, there are some problems···and they cannot be corrected in time."* [P02-L2]

> *"For some questions, I have trouble organizing a question. For example, a complicated question about having to transfer one thing to another unit. I may have chaos with word order or the language rules."* [P12-L2]

5.2.2 Pronouncing Unsure Terms Pronunciation of specific terms was also an issue when interacting with the IPAs across the devices used. Below are fragments from scenarios where participants were asked to get the IPAs to play music for them. Here, we see participants complete a scenario in which they are asked to play music by the artist Katy Perry.

Fragment 9

```
P03-L2 hey google ((murmur)) please:: open:: (.) katty::
       porry ↑music
GA     okay (.) check out this katy perry music station on youtube
```

Fragment 10

```
P06-L2 hey google (.) play the song of (.) kanty perry.
GH     i can't do that here, but you can ask me to play it on one
       of your other devices.
...
P06-L2 hey google (.) play the (0.4) music of katty perry
GH     ok (.) check out this youtube music station based on
       katy perry
```

(some)

In both of these fragments, the participants struggle to pronounce the artist's name, "Katy Perry". This is seen in the long pause before producing the name along with intonation (Fragment 9) suggesting that the participant is unsure whether their pronunciation will be recognised by the IPA. In Fragment 10, the participant seems to diagnose the issue not as one of network connectivity, but pronunciation, hence the change in how they pronounced the singer's name and the added emphasis through a change in intonation. In both fragments, the IPA manages to complete the query regardless

of the pronunciation and intonation, highlighting some capability to respond to various (mis)pronunciations of words.

As seen in the fragment below, when confronted with pronunciation challenges, L2 speakers look for pronunciation support from the IPA:

Fragment 11

```
P09-L2 how to say the word B-A-L-B-A-O
GH     sorry (.) i can't help with that yet
P09-L2 ah you can't help ((murmur))
       hey google how to:: (0.8) pronounce (1.1) the word
GH     that's pronounced cil bao
P09-L2 cil bao balbo ((murmur))
```

Issues in pronunciation are common in L2 speech, with the phonological set within a speaker's native languages having an influence on phonological encoding as well as pronunciation [25, 29, 46]. The fragments above highlight that, although this may not always impact the success of the interaction, they should still be considered in the language models used in speech recognition for L2 speech. Our fragments also show an opportunity for IPAs to offer pronunciation support to L2 speakers if they detect user difficulties. Post-interaction interviews further highlighted that L2 speakers tended to note how they placed significant emphasis on pronunciation when interacting with the IPAs, attempting to produce accurate pronunciation when generating commands to ensure that the IPA will interpret their command correctly:

"I tend to try to pronounce every single word accurately. Because I'm afraid the machines cannot understand me." [P11-L2]

It appears that, almost by default, L2 speakers tended to interpret errors as being due to their own pronunciation, despite our data also suggesting that IPAs can indeed handle some mispronunciations. When errors occurred, L2 speakers also mentioned finding it hard to come up with alternative strategies to ensure they were understood:

"...when you say the word, you think that you are pronouncing correctly. But actually it's not. The answer of the speech interface is wrong, and you don't know how to correct it." [P01-L2]

As evidenced in our interaction data, when pronunciation challenges occurred, L2 speakers tended to look to the IPA for support, wanting the IPA to help them:

"sometimes it cannot understand some pronunciation by non-native English speakers, and it cannot help you pronouncing the words you don't know." [P09-L2]

"As a non-native English speaker, I have a hard time on proper nouns···like Mozart. I have no idea about how to pronounce them correctly. English is my second language there are still lots of words, words like Celsius and Fahrenheit (which you don't know how to pronounce these words). If you want to ask a question about these words, you totally have no idea about how to ask this question...I don't know how to deal with that" [P04-L2]

5.2.3 Code Mixing Code mixing refers to the phenomenon seen in bilingual and multilingual speakers where linguistic units such as words and phrases from two or more languages appear within their utterances [47]. Within our data, we see that when interacting with the IPAs, L2 speakers engaged in code mixing, with words from their native language being combined with their L2 speech. As shown in the fragments below, this occurred both within commands as well as when participants were not addressing the IPA. The following fragment is of a participant looking to complete the scenario whereby they were asked to play the music of Mozart through Google Assistant on a smartphone. They are ultimately unsuccessful, using the Mandarin construction for the artist *Mòzhātè*.

Fragment 12

```
P04-L2 hey google (.) please open the mu- (0.3) eh musician (0.2)
       mòzh□tè
GA     here's some details
```

Participants also used their first language when talking to themselves between generating content for commands. Below we see an L2 participant murmur to themselves in Mandarin as they are trying to figure out how to request that the volume be turned down by the IPA (we have provided the translation on the next line).

Fragment 13

```
P03-L2 please (1.0) please turn down (1.3) ° 这是啥呀  °
                                           °what's that°
       ple- (.) plea (.) please turn down the voce
```

This type of effect is common when looking at L2 speech production. When struggling with L2 generation, L2 speakers are known to accidentally insert words from their L1 when generating utterances [38]. Difficulties in L1 lexical inhibition during production may lead to the types of code mixing common in L2 speech. Our fragments show that, not only does this happen in human-human dialogue, but also within IPA-based interactions.

5.2.4 Producing Preparatory Commands We also observed strategies practised exclusively by L2 participants, which supported the construction of commands. One prominent strategy involved the rehearsal of particular keywords, where L2 participants would rehearse their pronunciation or the structure of a command they wished to use. This was particularly common when participants encountered misunderstandings by the system. An example of this is shown in the fragment below, again from the scenario in which participants must convert a temperature from degrees Celsius to Fahrenheit.

Fragment 14

```
P08-L2 hey google (.) could you please help me to change the
       thirty two centis degree to another unit of the:: (.) eh
       temperature
GA     let's take a look
P08-L2 okay (.) ((murmur)) convert (.) fahrenheit (.) fahrenheit
       so:: (.) google (.) hey google (.) could you help me to
       change the thirty two centis degree to the fahrenheit
GA     let's take a look
```

As previously encountered in Section 5.2.1, the participant is attempting to convert the unit of temperature without specifically naming the intended unit. Following the first response by the IPA, the participant rehearses the pronunciation to themselves. This rehearsal is lower in volume and is not intended for the system to 'hear'. Yet, even with this rehearsal and the accurate pronunciation of Fahrenheit, the IPA still struggles to correctly identify the response. As P104-L2 states, however, completing commands without an understanding of which words to use or how to utter them is a significant barrier:

"English is your second language there are still lots of words, words like Celsius and Fahrenheit [...] if you want to ask a question about these words, you totally have no idea about how to ask this question." [P104-L2]

Below, we also see an L2 participant respond to this challenge by breaking up the same scenario into multiple commands, using the system to give them the information they need to complete the scenario in the study:

Fragment 15

```
P09-L2 hey google (.) there are two systems to:: describe the
       temperature (0.7) one:: is degree (.) how about the other
GA     according to dictionary.com: the two most common
       temperature scales (.) celsius and fahrenheit (.) are based
       on the freezing and boiling points of water (.) on the
       celsius scale there are one hundred increments between
       the two points (.) and on the fahrenheit scale there are
       one hundred and eighty
P09-L2 what's the temperature in dublin
       hey google (.) hey google? what's the temperature in dublin
GA     in dublin (.) it's currently six degrees
P09-L2 hey google (.) I want to change (0.4) thirty two degree
       to:: (0.8) another:: (0.9) system
GA     i found this on the web
```

Here the participant looks to the IPA to give them the words they need to complete the task (*Celsius* and *Fahrenheit*). They then look to use an exploratory task that is relevant to the scenario to test whether the output of the system indicates that it has understood them, using the output term for temperature (degree) in their utterance. Although the IPA has given them the information required to complete the task, the participant still does not then use the term 'Fahrenheit', leading the interaction to be unsuccessful. Thus, while our findings do suggest that L2 speakers have differing challenges with using IPAs, we also note how they are able to make use of features of the technology to overcome some of these. Next we will discuss our findings, reflecting on the relevancy to both literature and the design of IPAs.

6 Discussion

IPAs such as Google Assistant are now commonly available across numerous devices. Recent efforts have been made to extend access to IPAs to non-English speaking users. Yet, despite greater coverage in terms of supporting more languages other than English, IPA functionality is not equal across all the languages that are currently supported (e.g., [37]). For instance, Mandarin is currently not fully supported across the range of Google Assistant-enabled devices [33]. This means that, if they wish to use IPAs, these users are forced to use a second language to access the IPA's full functionality. Our research contributes by directly observing how people who interact with IPAs using a non-native language compare to people who interact with IPAs using their native language. Building on previous work [3, 48, 51], our findings show that L1 speakers tend to pause and hesitate when uncertain about how to construct a command, and that they tend to engage in second saying, partial and full repetition, and hyper-articulation when encountering errors. Our data suggests that L2 speakers also tend to use these strategies in interaction with IPAs, though their need to do so is more acute. We also identify many effects in command production that are unique to L2 speakers. These include language production issues such as difficulty in generating lexical items and appropriate syntactic structure. This results in using long pauses and needing more time to support command generation, particularly when encountering challenges in pronunciation. During production, there was also evidence of code mixing by combining L2 and L1 terms, which resulted in English and Mandarin words being included in commands. When encountering difficulties in interaction, L2 speakers also used off-task rehearsal and tended to break tasks into sub-tasks to help in successfully generating their commands. These findings contribute to the field by giving important insight into how L2 speakers vary in their approach to interacting with IPAs across both smart speakers and smartphones. This could have significant impact on how IPAs should be designed to support users who are interacting in a second language. We explore the potential reasons for our findings and the impact these could have on future IPA design below.

6.1 Language production mechanisms & L2 command construction

As highlighted above, our work finds a number of L2-specific issues in command construction, such as increased pausing during command generation, code mixing, as well as difficulties in lexical retrieval and grammatical construction. Research from psycholinguistics can shed light on the reasons for such effects. For instance, research shows that L2 speakers tend to generate utterances in a non-native language serially, rather than in parallel, as is common in L1 speech [40, 41, 57], making the production process slower [25]. Additionally, bilingual speakers need to store words for both languages in their mental lexicon, increasing retrieval time [25, 32]. This complexity tends to lead L2 speakers to hesitate more when producing utterances in their non-native language [25, 44]. L2 speakers also have difficulty generating non-native language utterances when both L1 and L2 languages do not share similar phonological systems, lexicons, or grammatical rules [16, 25] [58]. When compared to L1 speakers, L2 speakers' mental lexicons are activated in tandem [15, 23, 25, 38], leading to potentially semantically related words being activated across both language sets [15, 16, 38]. This can cause errors in lexical selection, lexical competition from the L2 speaker's native language, as well as the need to inhibit lexical alternatives [16, 17]. Such theoretical insights explain the reason why we see specific patterns within L2 speech in IPA interaction, whereby L2 speakers need more time for command generation and also experience a higher mental workload when constructing commands [62]. Based on this, future work should explore more systematically how specific L2 language production mechanisms impact IPA interaction. This would add much needed theoretical insight [14] but also support theory-driven efforts to assist L2 speakers in their efforts to generate effective IPA commands.

Future work might also consider differences between different groups of L2 speakers. For instance, the amount of time needed to generate a command might be different for L2 speakers whose native language is etymologically closer to English. A native Dutch speaker, for example, might need less time to construct a command in English than a native Mandarin speaker, due to closer parallels between English and Dutch in terms of lexicon and syntax. In taking a more granular approach, future work should thus look at

comparing how command construction varies between L2 speakers with similar L1 language structures.

6.2 Designing to support L2 speaker command generation

Our findings show that L2 users can experience significant challenges in generating appropriate terms and linguistic structures within IPA command construction. We imagine a number of features that IPAs might include to better support L2 speakers. For instance, the development of voice activity detection algorithms that are more sensitive to the need of L2 speakers to pause and rehearse their utterances may positively benefit the L2 speaker user experience. Time allowed might be informed by personalised measures of L2 proficiency, or from the etymological relationship between languages. Production support could also be better achieved through the increased use of screen-based output on both smart speaker and smartphone IPAs. Screen-based feedback is particularly important for L2 speakers in IPA interaction as it supports users in diagnosing errors [63], which currently seem to be over-attributed to pronunciation issues. Screens could be used to suggest alternative phrasings or prime users with terms or command structures that are likely to bring the most success in interaction. Both lexical and syntactic priming [10, 18] are common in human-machine dialogue. These types of effects could be leveraged to support L2 production.

Furthermore, the knowledge that L2 speakers engage in code mixing when interacting with IPAs could be used to inform the development of flexible speech recognition methods, whereby terms from different languages are understood and incorporated into the commands recognised. To better support L2 speakers' difficulties around pronunciation, a more supportive IPA could allow users to get feedback when pronunciation leads to errors. For example, IPAs saying something like "Here's what I heard:" when ASR accuracy probability predictions are below, say, 70% for example, or when users explicitly acknowledge a query has not been recognised accurately. This may help L2 speakers diagnose pronunciation errors more effectively and thus modify their pronunciation to accommodate the system only when required. IPA-driven pronunciation support in particular was emphasised within post-interaction interviews as potentially beneficial to the L2 user experience.

Finally, as different individuals use idiosyncratic strategies to support their unique language needs, allowing more customisable features, like user-specific lexicons (e.g., "When I say 'F' I mean Fahrenheit") may help better tailor IPAs to individuals, particularly in supporting routine commands or specific words that frequently lead to recognition errors. Future work within the HCI community should explore how these proposed features may influence command construction and the L2 user experience.

7 Limitations

The work presented aimed to identify the differing patterns in language production across L1 and L2 speakers when interacting with IPAs. Specifically, our work chose to focus on L2 speakers of English who held Mandarin as their native language. This was so the interviews, used to support our linguistic findings, could be conducted by the lead author in their native language to gain as much detail as possible regarding their experiences. We also chose Mandarin speakers as the language is vastly different linguistically from English, allowing us to make a clear distinction between the native languages for L1 and L2 speakers in our study. We recommend that future work looks to observe whether similar patterns occur with participants who have native languages other than Mandarin.

It is also important to note that all participants were students at an English-speaking European University. Therefore, all L2 speakers studied and lived in an English-speaking country, where they are likely to be regularly exposed to, and be proficient in, English. We believe that this may make our findings a conservative indicator of the issues and patterns of L2 speakers, as those who are less proficient or less exposed to English may exhibit similar behaviours, but experience even greater difficulty when interacting with the IPA. Although we explore the language patterns from an observational perspective, it is also important to note that these were generated during a lab-based study. We used a lab-based study as this meant we could minimise the influence of background noise and other distractions on task performance. It also allowed us to control the tasks that people were asked to complete, ensuring that we could identify the aim of the participants and when they had completed the task.

As part of the experiment, we asked participants to complete two sets of six scenarios that were delivered as pictograms, rather than using text- or audio-based instructions. Using pictograms was chosen to ensure that participants could not directly copy any written or audio scenario descriptions when generating their command, whilst keeping the command generation process closer to the experience of 'real world' IPA command generation. The use of pictograms also meant that both sets of scenarios could be used without the need for text or audio translation for L2 participants. To ensure that all pictograms were interpreted correctly and to minimise learning effects during the experiment scenarios, we asked all participants to report how they would word a query to complete a similar practice set of scenario pictograms, delivered before IPA interaction. This gave participants the opportunity to clarify any interpretation issues before the study session commenced. Future work should be cognisant of the impact of task delivery and task interference on IPA command construction.

8 Conclusion

Intelligent Personal Assistants (IPAs) currently do not support all languages, meaning many users often have to use a second language to interact with them. Our work focuses on how L2 users interact with IPAs and how this compares to L1 speakers, with a particular focus on the generation of commands. Our Conversation Analysis of the interaction transcripts along with our interview data suggests there are similarities in how L1 and L2 speakers approach command generation. Both use pauses when unsure about pronunciation or how to form a sentence, and both use partial or complete repetition of commands and/or hyper-articulation when encountering errors. We also found L2-specific effects, which focused on issues in language production and command generation, including lexical access, syntactic construction, and pronunciation, as well as the use of code mixing. L2 speakers in particular

also used strategies such as increased pause lengths and off-task rehearsal to help generate commands, with varying success. Based on these, we suggest IPAs should be designed to be more sensitive to the interaction needs of L2 speakers, allowing IPAs to become more inclusive technologies.

Acknowledgments

This work was conducted with the financial support of the UCD China Scholarship Council (CSC) Scheme grant No. 201908300016, Science Foundation Ireland ADAPT Centre under Grant No. 13/RC/2106 and the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (D-REAL) under Grant No. 18/CRT/6224.

References

[1] Saul Albert and Magnus Hamann. 2021. Putting Wake Words to Bed: We Speak Wake Words with Systematically Varied Prosody, but CUIs Don't Listen. In *CUI 2021 - 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 13, 5 pages. https://doi.org/10.1145/3469595.3469608

[2] René Amalberti, Noëlle Carbonell, and Pierre Falzon. 1993. User representations of computer systems in human-computer speech interaction. *International Journal of Man-Machine Studies* 38, 4 (April 1993), 547–566. https://doi.org/10.1006/imms.1993.1026

[3] René Amalberti, Noëlle Carbonell, and Pierre Falzon. 1993. User representations of computer systems in human-computer speech interaction. *International Journal of Man-Machine Studies* 38, 4 (April 1993), 547–566. https://doi.org/10.1006/imms.1993.1026

[4] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput.-Hum. Interact.* 26, 3, Article 17 (April 2019), 28 pages. https://doi.org/10.1145/3311956

[5] Sungeun An, Robert Moore, Eric Young Liu, and Guang-Jie Ren. 2021. Recipient Design for Conversational Agents: Tailoring Agent's Utterance to User's Knowledge. In *CUI 2021 - 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, Article 30, 5 pages. https://doi.org/10.1145/3469595.3469625

[6] J Maxwell Atkinson and John Heritage. 1984. Transcript Notation. In *Structures of Social Action: Studies in Conversation Analysis*. Cambridge University Press, Cambridge, UK, ix–xvi. https://doi.org/10.1017/CBO9780511665868

[7] Liam Bannon, John Bowers, Peter Carstensen, John A Hughes, Kari Kuutii, James Pycock, Tom Rodden, Kjeld Schmidt, Dan Shapiro, Wes Sharrock, and Stephen Viller. 1993. Informing CSCW System Requirements. In *COMIC Deliverable 2.1*. The COMIC Project, Lancaster, UK.

[8] Linda Bell and Joakim Gustafson. 1999. Interaction with an animated agent in a spoken dialogue system. In *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH '99)*. ESCA, Grenoble, France, 1143–1146.

[9] Holly P Branigan, Martin J Pickering, Jamie Pearson, Janet F McLean, and Ash Brown. 2011. The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition* 121, 1 (2011), 41–57.

[10] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, Janet F. McLean, and Ash Brown. 2011. The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition* 121, 1 (Oct. 2011), 41–57. https://doi.org/10.1016/j.cognition.2011.05.011

[11] Susan E Brennan. 1998. The Grounding Problem in Conversations With and Through Computers. In *Social and cognitive approaches to interpersonal communication*. Psychology Press, New York, NY, USA, Chapter 9, 201–225.

[12] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* 31, 4 (09 2019), 349–371. https://doi.org/10.1093/iwc/iwz016 arXiv:https://academic.oup.com/iwc/article-pdf/31/4/349/33525046/iwz016.pdf

[13] Leigh Clark, Abdulmalik Ofemile, and Benjamin R. Cowan. 2021. Exploring Verbal Uncanny Valley Effects with Vague Language in Computer Speech. In *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers*, Benjamin Weiss, Jürgen Trouvain, Melissa Barkat-Defradas, and John J. Ohala (Eds.). Springer Singapore, Singapore, 317–330. https://doi.org/10.1007/978-981-15-6627-1_17

[14] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300705

[15] Albert Costa and Alfonso Caramazza. 1999. Is lexical selection in bilingual speech production language-specific? Further evidence from Spanish–English and English–Spanish bilinguals. *Bilingualism: Language and cognition* 2, 3 (1999), 231–244.

[16] Albert Costa, Michele Miozzo, and Alfonso Caramazza. 1999. Lexical selection in bilinguals: Do words in the bilingual's two lexicons compete for selection? *Journal of Memory and language* 41, 3 (1999), 365–397.

[17] Albert Costa, Mikel Santesteban, and Iva Ivanova. 2006. How do highly proficient bilinguals control their lexicalization process? Inhibitory and language-specific selection mechanisms are both functional. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32, 5 (2006), 1057.

[18] Benjamin R Cowan, Holly P Branigan, Mateo Obregón, Enas Bugis, and Russell Beale. 2015. Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human- computer dialogue. *International Journal of Human-Computer Studies* 83 (2015), 27–42.

[19] Benjamin R. Cowan, Philip Doyle, Justin Edwards, Diego Garaialde, Ali Hayes-Brady, Holly P. Branigan, João Cabral, and Leigh Clark. 2019. What's in an Accent? The Impact of Accented Synthetic Speech on Lexical Choice in Human-Machine Dialogue. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 23, 8 pages. https://doi.org/10.1145/3342775.3342786

[20] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What Can i Help You with?": Infrequent Users' Experiences of Intelligent Personal Assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 43, 12 pages. https://doi.org/10.1145/3098279.3098539

[21] Anna L. Cox, Paul A. Cairns, Alison Walton, and Sasha Lee. 2008. Tlk or txt? Using voice input for SMS composition. *Personal and Ubiquitous Computing* 12, 8 (2008), 567–588. https://doi.org/10.1007/s00779-007-0178-8

[22] Andy Crabtree, Peter Tolmie, and Mark Rouncefield. 2013. "How Many Bloody Examples Do You Want?" Fieldwork and Generalisation. In *ECSCW 2013: Proceedings of the 13th European Conference on Computer Supported Cooperative Work, 21-25 September 2013, Paphos, Cyprus*, Olav W. Bertelsen, Luigina Ciolfi, Maria Antonietta Grasso, and George Angelos Papadopoulos (Eds.). Springer, London, 1–20. https://doi.org/10.1007/978-1-4471-5346-7_1

[23] Kees De Bot. 2007. A Bilingual Production Model: Levelt's 'Speaking' Model Adapted. In *The Bilingualism Reader*. Routledge, London, UK, 384–404.

[24] Nivja H De Jong. 2016. Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching* 54, 2 (2016), 113–132.

[25] Zoltán Dörnyei and Judit Kormos. 1998. Problem-solving mechanisms in L2 communication: A psycholinguistic perspective. *Studies in second language acquisition* 20, 3 (1998), 349–385.

[26] Philip R Doyle, Leigh Clark, and Benjamin R. Cowan. 2021. What Do We See in Them? Identifying Dimensions of Partner Models for Speech Interfaces Using a Psycholexical Approach. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 244, 14 pages. https://doi.org/10.1145/3411764.3445206

[27] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI ' 19)*. Association for Computing Machinery, New York, NY, USA, Article 5, 12 pages. https://doi.org/10.1145/3338286.3340116

[28] Mateusz Dubiel, Martin Halvey, and Leif Azzopardi. 2018. A Survey Investigating Usage of Virtual Personal Assistants. *CoRR* abs/1807.04606 (2018), 5 pages. arXiv:1807.04606 http://arxiv.org/abs/1807.04606

[29] Claus Faerch and Gabriele Kasper. 1986. Strategic Competence in Foreign Language Teaching. In *Learning, teaching and communication in the foreign language classroom*. Aarhus University Press, Aarhus, Denmark, 179–193.

[30] Joel E. Fischer, Stuart Reeves, Martin Porcheron, and Rein Ove Sikveland. 2019. Progressivity for Voice Interface Design. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI ' 19)*. Association for Computing Machinery, New York, NY, USA, Article 26, 8 pages. https://doi.org/10.1145/3342775.3342788

[31] Emer Gilmartin, Marine Collery, Ketong Su, Yuyun Huang, Christy Elias, Benjamin R. Cowan, and Nick Campbell. 2017. Social talk: making conversation with people and machine. In *Proceedings of the 1st Association for Computing Machinery SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents - ISIAA 2017*. ACM Press, Glasgow, UK, 31–32. https://doi.org/10.1145/3139491.3139494

[32] Tamar H Gollan and Lori-Ann R Acenas. 2004. What is a TOT? Cognate and translation effects on tip-of-the-tongue states in Spanish-English and tagalog-English bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30, 1 (2004), 246.

[33] Google, Inc. 2021. Talk to the Google Assistant in multiple languages. Retrieved Sep 9, 2021 from https://support.google.com/googlenest/answer/7550584?hl=en-AU&co=GENIE.Platform%3DAndroid#zippy=%2Cgoogle-nest-hub-max%2Cgoogle-nest-hub

[34] Eric Hauser. 2017. Learning and the Immediate Use(fulness) of a New Vocabulary Item. *The Modern Language Journal* 101, 4 (2017), 17. https://doi.org/10.1111/modl.12429

[35] Gail Jefferson et al. 2004. Glossary of transcript symbols with an introduction. *Pragmatics and beyond new series* 125 (2004), 13–34.

[36] Alan Kennedy, A Wilkes, L Elder, and Wayne Murray. 1988. Dialogue with machines. *Cognition* 30 (1988), 37–72. https://doi.org/10.1016/0010-0277(88)90003-0

[37] Bret Kinsella. 2019. Google Assistant Now Supports Simplified Chinese on Android Smartphones. http://bit.ly/30Yg8qN. Accessed 27th Jan 2020.

[38] Wido La Heij. 2005. Selection Processes in Monolingual and Bilingual Lexical Access. In *Handbook of Bilingualism: Psycholinguistic Approaches*, Judith F Kroll and A M B de Groot (Eds.). Oxford University Press, Oxford, UK.

[39] Ludovic Le Bigot, Jean-François Rouet, and Eric Jamet. 2007. Effects of Speech- and Text-Based Interaction Modes in Natural Language Human-Computer Dialogue. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 49, 6 (Dec. 2007), 1045–1053. https://doi.org/10.1518/001872007X249901

[40] Willem JM Levelt, Ardi Roelofs, and Antje S Meyer. 1999. A theory of lexical access in speech production. *Behavioral and brain sciences* 22, 1 (1999), 1–38.

[41] Willem JM Levelt and Herbert Schriefers. 1987. Stages of lexical access. In *Natural language generation*. Springer, Dordrecht, Netherlands, 395–404.

[42] Irene Lopatovska, Katrina Rink, Ian Knight, Kieran Raines, Kevin Cosenza, Harriet Williams, Perachya Sorsche, David Hirsch, Qi Li, and Adrianna Martinez. 2019. Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science* 51, 4 (2019), 984–997.

[43] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[44] Dorothea Möhle. 1984. A comparison of the second language speech production of different native speakers. *Second language productions* 26 (1984), 49.

[45] Robert J Moore, Margaret H Szymanski, and Raphael Arar (Eds.). 2018. *Studies in Conversational UX Design* (1 ed.). Springer International Publishing, Cham, Switzerland. 204 pages. https://doi.org/10.1007/978-3-319-95579-7

[46] Pieter Muysken. 2013. Two Linguistic Systems in Contact: Grammar, Phonology, and Lexicon. In *The handbook of bilingualism and multilingualism*. Blackwell Publishing, Ltd, Chichester, UK, Chapter 8, 193–216. https://doi.org/10.1002/9781118332382.ch8

[47] Pieter Muysken, Pieter Cornelis Muysken, et al. 2000. *Bilingual speech: A typology of code-mixing*. Cambridge University Press.

[48] Sharon Oviatt, Jon Bernard, and Gina-Anne Levow. 1998. Linguistic Adaptations During Spoken and Multimodal Error Resolution. *Language and Speech* 41, 3-4 (July 1998), 419–442. https://doi.org/10.1177/002383099804100409

[49] Hannah R M Pelikan and Mathias Broth. 2016. Why That Nao?: How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4921–4932. https://doi.org/10.1145/2858036.2858478

[50] Martin Porcheron. 2021. What's in a name and does CUI matter?. In *CUI 2021 - 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) *(CUI '21)*. Association for Computing Machinery, New York, NY, USA, 3 pages. https://doi.org/10.1145/3469595.3469619

[51] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 ACM Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article 640, 12 pages. https://doi.org/10.1145/3173574.3174214

[52] Martin Porcheron, Joel E Fischer, and Sarah Sharples. 2017. "Do Animals Have Accents?": Talking with Agents in Multi-Party Conversation. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 207–219. https://doi.org/10.1145/2998181.2998298

[53] Aung Pyae and Paul Scifleet. 2018. Investigating Differences between Native English and Non-Native English Speakers in Interacting with a Voice User Interface: A Case of Google Home. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction* (Melbourne, Australia) *(OzCHI ' 18)*. Association for Computing Machinery, New York, NY, USA, 548–553. https://doi.org/10.1145/3292147.3292236

[54] Aung Pyae and Paul Scifleet. 2019. Investigating the Role of User's English Language Proficiency in Using a Voice User Interface: A Case of Google Home Smart Speaker. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA ' 19)*. Association for Computing Machinery, New York, NY, USA, 6 pages. https://doi.org/10.1145/3290607.3313038

[55] Stuart Reeves and Martin Porcheron. 2022. Conversational AI: Respecifying Participation as Regulation. In *The Sage Handbook of Digital Society*, William Housley, Adam Edwards, Roser Montagut, and Richard Fitzgerald (Eds.). SAGE Publications.

[56] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language* 50, 4 (01 1974), 696–735. https://doi.org/10.2307/412243

[57] Norman Segalowitz and Jan Hulstijn. 2005. Automaticity in Bilingualism and Second Language Learning. In *Handbook of Bilingualism: Psycholinguistic Approaches*, Judith F Kroll and A M B de Groot (Eds.). Oxford University Press, Oxford, UK, 371–388.

[58] Catherine Watson, Wei Liu, and Bruce MacDonald. 2013. The effect of age and native speaker status on synthetic speech intelligibility. In *Proceedings of the 8th ISCA Workshop on Speech Synthesis (SSW 8)*. ISCA, Grenoble, France, 195–200.

[59] Richard Wiese. 1984. Language Production in Foreign and Native Languages: Same or Different? In *Second Language Productions*, Hans W Dechert, Dorothea Möhle, and Manfred Raupach (Eds.). Narr Verlag, Tübingen, Germany, 11–25.

[60] Jean Wong. 2000. Repetition in Conversation: A Look at "First and Second Sayings". *Research on Language and Social Interaction* 33, 4 (2000), 407–424. https://doi.org/10.1207/S15327973RLSI3304_03

[61] Robin Wooffitt. 1990. On the Analysis of Interaction: An Introduction to Conversation Analysis. In *Computers and Conversation*, Paul Luff, David Frohlich, and Nigel Gilbert (Eds.). Academic Press, San Diego, CA, USA, Chapter 1, 7–38.

[62] Yunhan Wu, Justin Edwards, Orla Cooney, Anna Bleakley, Philip R. Doyle, Leigh Clark, Daniel Rough, and Benjamin R. Cowan. 2020. Mental Workload and Language Production in Non-Native Speaker IPA Interaction. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 3, 8 pages. https://doi.org/10.1145/3405755.3406118

[63] Yunhan Wu, Daniel Rough, Anna Bleakley, Justin Edwards, Orla Cooney, Philip R. Doyle, Leigh Clark, and Benjamin R. Cowan. 2020. See What I'm Saying? Comparing Intelligent Personal Assistant Use for Native and Non-Native Language Speakers. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services* (Oldenburg, Germany) *(MobileHCI '20)*. Association for Computing Machinery, New York, NY, USA, Article 34, 9 pages. https://doi.org/10.1145/3379503.3403563

[64] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. 2017. Achieving Human Parity in Conversational Speech Recognition. arXiv:1610.05256 [cs.CL]