Evaluating automatic speech recognition for L2 pronunciation feedback: a focus on Google Translate

Paul John[1], Walcir Cardoso[2], and Carol Johnson[3]

Abstract. This study examines the L2 pronunciation feedback provided by the Automatic Speech Recognition (ASR) functionality in Google Translate (GT). We focus on three Quebec Francophone (QF) errors in English: th-substitution, h-deletion, and h-epenthesis. Four hundred and eighty male and female QF recordings of sentences with correctly and incorrectly pronounced final items (e.g. *I don't know who to thank* versus *tank*) were played into GT. Errors were equally divided between mispronunciations leading to real word (*thank* → *tank*) and nonword output (*thief* → *tief*). As anticipated, we found greater transcription accuracy for correct pronunciations and, among incorrect pronunciations, for real words versus nonwords. Overall, our findings suggest ASR can be highly effective for pronunciation feedback. We also examined transcriptions for gender bias, since ASR systems are often trained on corpora with more male voices, but our concerns proved unfounded: surprisingly, higher transcription accuracy was found for female recordings.

Keywords: automatic speech recognition, Google Translate, L2 pronunciation, corrective feedback, gender bias.

1. Introduction

ASR technology constitutes a promising means for second language (L2) learners to access feedback on pronunciation errors. For example, QFs typically struggle with English /θ/ and /h/, tending to substitute /t/ for /θ/ (*thank* → *tank*), and to delete or epenthesize /h/ (*heat* → *_eat* / *air* → *hair* respectively) (Brannen, 2011; John & Cardoso, 2009). If QFs use ASR to transcribe their output for targets

1. Université du Québec à Trois-Rivières, Trois-Rivières, Canada; paul.john@uqtr.ca; https://orcid.org/0000-0002-6963-7988
2. Concordia University, Montreal, Canada; walcir.cardoso@concordia.ca; https://orcid.org/0000-0001-6376-185X
3. Concordia University, Montreal, Canada; carol.johnson@concordia.ca; https://orcid.org/0000-0001-6859-6566

How to cite this article: John, P., Cardoso, W., & Johnson, C. (2022). Evaluating automatic speech recognition for L2 pronunciation feedback: a focus on Google Translate. In B. Arnbjörnsdóttir, B. Bédi, L. Bradley, K. Friðriksdóttir, H. Garðarsdóttir, S. Thouësny, & M. J. Whelpton (Eds), *Intelligent CALL, granular systems, and learner data: short papers from EUROCALL 2022* (pp. 197-202). Research-publishing.net. https://doi.org/10.14705/rpnet.2022.61.1458

© 2022 Paul John, Walcir Cardoso, and Carol Johnson (CC BY)

such as *thank, heat*, and *air* (e.g. in a reading-aloud task), they may find that the transcription reflects the incorrect pronunciations *tank, eat*, and *hair*. The transcription thus provides the invaluable feedback that learners have produced instances of th-substitution, h-deletion, and h-epenthesis, and learners can revise their pronunciation until the transcription matches the target.

Corrective feedback is an effective means of promoting L2 pronunciation accuracy (Lyster, Saito, & Sato, 2013). Immediate feedback is, however, hard to provide, and delayed feedback (e.g. on recordings) is time-consuming for teachers to formulate, so learners may not receive much feedback. This is where ASR can fill the gap, generating feedback that learners access 'anytime-anywhere' to engage in autonomous learning (van Lieshout & Cardoso, 2022). Nonetheless, questions remain regarding the adequacy of ASR-generated feedback, as in the widely available tool GT. To what extent do GT transcriptions capture correct and incorrect pronunciation?

Our study investigates transcription accuracy for items appearing in sentence contexts rather than in isolation (e.g. in wordlists). In this case, GT can identify the target item using not only phonetic but also contextual (syntactic/collocational/semantic) cues (Ashwell & Elam, 2017). Under these conditions, we expected higher transcription accuracy for correctly than incorrectly pronounced target items. Given correct pronunciation (*I don't know who to thank*), phonetic and contextual cues converge on the target item. Given incorrect pronunciation (*I don't know who to tank*), phonetic and contextual cues conflict, and GT may transcribe contextually motivated *thank* rather than phonetically accurate *tank*, thereby failing to flag the pronunciation error. The adequacy of ASR-based feedback depends on transcriptions both confirming correct and flagging incorrect pronunciations.

With incorrect pronunciations, we also anticipated greater accuracy for real words (*thank → tank*; *heat → _eat*; *air → hair*) versus nonword output (*thief → tief; head → _ead; ice → hice*). Nonwords being absent from the GT lexicon, the technology should fail to flag such errors, often supplying the contextually appropriate item. Finally, we investigated ASR for gender bias: since ASR is often trained on corpora with more male voices (Tatman, 2017), we anticipated a male transcription advantage.

The following summarizes our predictions regarding transcription accuracy (with '>' indicating 'greater than'): *correct > incorrect pronunciations*; *real word > nonword output*; *male > female speech*.

2. Method

One hundred and twenty sentences were recorded by four male (M) and four female (F) QF adults with correct and incorrect pronunciation of final items starting with /θ/, /h/, or a vowel. Among incorrect pronunciations, 60 led to real word and 60 to nonword output. Four hundred and eighty recordings (four versions of each sentence: 1M/1F recording with correct/incorrect pronunciation) were played into GT and coded for final-item transcription accuracy.

Among inaccurate transcriptions for correctly/incorrectly pronounced items, we also determined rates of 'false alarms' and 'false negatives'. A false alarm involves, for example, correctly realized *thank* being transcribed as *tank,* misleadingly suggesting the learner has substituted /t/ for /θ/. A false negative involves *thank* being transcribed as *thank,* despite being incorrectly realized as *tank,* misleadingly indicating target-like pronunciation.

3. Results and discussion

Table 1 presents accuracy rates for transcriptions of correctly pronounced sentence-final items. The overall accuracy rate (88.33%) for correct pronunciations is reassuring: in most cases, GT confirmed the target-like pronunciation.

Furthermore, among inaccurate transcriptions (11.66%), we can report that fully half (5.83%) constitute 'near-accurate' transcriptions. That is, the mistranscription nonetheless started with the problematic target sound (e.g. output *thrifty* was transcribed as *thirsty*), from which learners can correctly conclude that they successfully realized the target sound (/θ/ in this example). In addition, no false alarms occurred among the mistranscriptions: correctly pronounced *thank-heat-air* were never transcribed as *tank-eat-hair*. Again, the absence of such misleading feedback is encouraging.

Table 1. Transcription accuracy: correct pronunciations (%)

Target items	**M**	**F**	**M + F**
th-initial	72.50	85.00	78.75
h-initial	90.00	97.50	93.75
V-initial	87.50	97.50	92.50
Mean	**83.33**	**93.33**	**88.33**

Table 2 presents accuracy rates for transcriptions of incorrectly pronounced sentence-final items leading to real word (a) and nonword output (b).

Table 2. Transcription accuracy: incorrect pronunciations (%)

a. Real word output (thank → tank)			
Target items	**M**	**F**	**M + F**
th-initial	30.00	40.00	35.00
h-initial	35.00	65.00	50.00
V-initial	30.00	85.00	57.50
Mean	**31.66**	**63.33**	**47.50**
b. Nonword output (thief → tief)			
Target items	**M**	**F**	**M + F**
th-initial	0.00	0.00	0.00
h-initial	5.00	15.00	10.00
V-initial	10.00	20.00	15.00
Mean	**5.00**	**11.66**	**8.33**

As expected, the overall mean for incorrect pronunciations resulting in real words (47.5%) is considerably lower than for correct pronunciations (88.33%). The mean for nonword output (8.33%) is lower still, so GT is virtually incapable of providing feedback on nonword mispronunciations. Exceptions were mainly due to the technology identifying proper nouns corresponding to phonetic input (e.g. *oil* → *hoil* was transcribed as *Hoyle*, a place name). This is important information for teachers who wish to design ASR-based pronunciation activities: these should target items that, if mispronounced, lead to real word output. Since GT flags almost half the real word errors here, it constitutes an effective tool for providing QF ESL learners with pronunciation feedback. Moreover, some mistranscriptions (14.5% in all) were 'near-accurate': for example, output *teft* for *theft* was transcribed as *test*, which captures the mispronunciation.

Nonetheless, GT generated numerous false negatives: 36.66% (real word output) and 65% (nonword output). Consequently, to verify pronunciation ability, learners should test themselves on multiple tokens, practicing until transcriptions are consistently accurate. Alternatively, learners can produce items in isolation, thus eliminating the possibility that GT bases its transcription on contextual cues.

Interestingly, contra the expected pattern for gender bias, transcription accuracy for F recordings is higher than for M recordings across the board, whether for correct (Table 1) or incorrect pronunciations (Table 2). Conceivably, the F recordings contained the more careful pronunciation that typifies female L2 speech (Moyer,

2016). This hypothetically clearer articulation may facilitate automatic recognition of individual items and override any inherent gender bias in the technology.

4. Conclusions

Our findings indicate that ASR is a highly promising tool for much-needed L2 pronunciation feedback. GT showed high transcription accuracy for correct pronunciations and no false alarms. Moreover, the consistently higher transcription accuracy for female voices suggests that concerns about ASR gender bias are unfounded. Although ASR struggles with nonword errors (8.33% transcription accuracy), almost half of real word errors in a sentence context were flagged, and we would expect even higher transcription accuracy for items spoken in isolation. Indeed, the next stage of research will target wordlists, such that transcriptions are based on phonetic input alone, without contextual cues. Future research could also go beyond consonants to include L2 vowel and lexical stress errors. In sum, while the technology has certain limitations for teachers to highlight and explain, L2 learners can use ASR to access invaluable feedback on pronunciation.

5. Acknowledgments

This project received funding from the *Luc Maurice Foundation* and the *Social Sciences and Humanities Research Council* (Canada).

References

Ashwell, T., & Elam, J. (2017). How accurately can the Google Web Speech API recognize and transcribe Japanese L2 English learners' oral production? *The JALT CALL Journal, 13*(1), 59-76. https://doi.org/10.29140/jaltcall.v13n1.212

Brannen, K. (2011). *The perception and production of interdental fricatives in second language acquisition*. Unpublished doctoral dissertation, McGill University.

John, P., & Cardoso, W. (2009). Francophone ESL learners' difficulties with English /h/. In M. A. Watkins, A. S. Rauber, & B. O. Baptista (Eds), *Recent research in second language phonetics/phonology: perception and production* (pp. 118-140). Cambridge Scholars Publishing.

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching, 46*(1), 1-40. https://doi.org/10.1017/S0261444812000365

Moyer, A. (2016). The puzzle of gender effects in L2 phonology. *Journal of Second Language Pronunciation, 2*(1), 8-28. https://doi.org/10.1075/jslp.2.1.01moy

Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. *Proceedings of the First Workshop on Ethics in Natural Language Processing* (pp. 53-59). https://doi.org/10.18653/v1/W17-1606

Van Lieshout, C., & Cardoso, W. (2022). Google Translate as a tool for self-directed language learning. *Language Learning & Technology*, *26*(1), 1-19. http://hdl.handle.net/10125/73460

Published by Research-publishing.net, a not-for-profit association
Contact: info@research-publishing.net

© 2022 by Editors (collective work)
© 2022 by Authors (individual work)

Intelligent CALL, granular systems and learner data: short papers from EUROCALL 2022
Edited by Birna Arnbjörnsdóttir, Branislav Bédi, Linda Bradley, Kolbrún Friðriksdóttir, Hólmfríður Garðarsdóttir, Sylvie Thouësny, and Matthew James Whelpton

Publication date: 2022/12/12

Rights: the whole volume is published under the Attribution-NonCommercial-NoDerivatives International (CC BY-NC-ND) licence; **individual articles may have a different licence**. Under the CC BY-NC-ND licence, the volume is freely available online (https://doi.org/10.14705/rpnet.2022.61.9782383720157) for anybody to read, download, copy, and redistribute provided that the author(s), editorial team, and publisher are properly cited. Commercial use and derivative works are, however, not permitted.

Disclaimer: Research-publishing.net does not take any responsibility for the content of the pages written by the authors of this book. The authors have recognised that the work described was not published before, or that it was not under consideration for publication elsewhere. While the information in this book is believed to be true and accurate on the date of its going to press, neither the editorial team nor the publisher can accept any legal responsibility for any errors or omissions. The publisher makes no warranty, expressed or implied, with respect to the material contained herein. While Research-publishing.net is committed to publishing works of integrity, the words are the authors' alone.

Trademark notice: product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Copyrighted material: every effort has been made by the editorial team to trace copyright holders and to obtain their permission for the use of copyrighted material in this book. In the event of errors or omissions, please notify the publisher of any corrections that will need to be incorporated in future editions of this book.

Typeset by Research-publishing.net
Cover photo by © 2022 Kristinn Ingvarsson (photo is taken inside Veröld – House of Vigdís)
Cover layout by © 2022 Raphaël Savina (raphael@savina.net)

ISBN13: 978-2-38372-015-7 (PDF, colour)

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2022.