Gallant, J. (2020). Typed transcription tasks: A simultaneous measure of the intelligibility and comprehensibility of individual words. In O. Kang, S. Staples, K. Yaw, & K. Hirschi (Eds.), *Proceedings of the 11th Pronunciation in Second Language Learning and Teaching conference*, ISSN 2380-9566, Northern Arizona University, September 2019 (pp. 86–96). Ames, IA: Iowa State University.

TYPED TRANSCRIPTION TASKS: A SIMULTANEOUS MEASURE OF THE INTELLIGIBILITY AND COMPREHENSIBILITY OF INDIVIDUAL WORDS

Jordan Gallant, Brock University

Typing transcription tasks offer a direct, on-line, simultaneous measure of intelligibility and comprehensibility. This study reports results from a typed transcription task of individual words comparing responses to Mandarin- and native English-accented speech. Analysis focuses on inter-keystroke latencies and proportion of corrected errors. Results show significantly greater individual keystroke latencies and lower proportions of corrected errors for Mandarin-accent stimuli. Potential underlying causes for these results are discussed. This study concludes that the data collected in typed transcription tasks provide deeper insights into our understanding of intelligibility and comprehensibility.

INTRODUCTION

Evaluating the communicative success of foreign-accented speech has been approached using several distinct, yet related, dimensions. This paper will focus on two of these dimensions: intelligibility and comprehensibility. Intelligibility refers to "the extent to which an utterance is understood" (Munro & Derwing, 1995, p. 291) and comprehensibility to a listener's "perceptions of difficulty in understanding particular utterances" (Munro & Derwing, 1995, p. 291). Intelligibility is most often operationalized through binary, right-wrong measures of whether an utterance is understood by the listener and comprehensibility through scalar ratings of perceived listener effort. The distinction between these dimensions is important because it recognizes that while two utterances may both be successfully understood, the amount of effort required to process them may differ significantly.

Comprehensibility corresponds to processes that occur in the mind of the listener during on-line speech processing. Framed this way, comprehensibility may be more transparently referred to as 'processability' (Thomson, 2018) or processing fluency, and therefore, optimally operationalized through direct, on-line behavioral measures, which reflect the processes occurring in the mind of the listener. Ideally, a task providing a direct, on-line measure of both intelligibility and comprehensibility would be ideal for evaluating the communicative success of foreign-accented speech. This paper presents results from a typed transcription task and discusses the advantages of this method in terms of data collection.

Typed Transcription Tasks

Typed transcription tasks (Libben & Weber, 2014) simply involve listeners' responses being typed on a keyboard, rather than written down using pen and paper. The advantage of typed responses is that every keystroke and its response latency can be collected and analyzed. This allows for more detailed error analysis, including analysis of on-line production errors which are subsequently corrected, as well as chronometric analyses of on-line typed responses. Thus, typing transcription tasks provide both a measure of intelligibility in terms of keystroke accuracy, and comprehensibility in terms of keystroke latency.

Operationalizing Intelligibility: Response Accuracy

Intelligibility has been chiefly operationalized through response accuracy in transcription/dictation tasks in which participants identify and transcribe auditory stimuli. Accuracy is calculated by comparing the proportion of responses correctly identified and transcribed. While many variations of transcription task appear in the literature (e.g., sentence transcription, cloze dictation, sentence completion), typed transcription tasks are compatible with all of the variants.

The collation of response accuracy in transcription tasks differs from study to study. While all tasks have an objectively correct answer, the criteria used in the collation of response accuracy can vary. For instance, misspellings of targets (Bent & Bradlow, 2003) or errors made in the transcription of function words in sentence transcription (Derwing and Munro, 1997) are often deemed trivial errors and collated as correct responses. These methods of response categorization fail to take in account differential trivial error rates across speakers. If intelligibility is the only dimension being measured, this is not an issue. However, it is conceivable that trivial errors are less likely to be noticed, and therefore, less likely to be corrected, when participants are involved in a more cognitively effortful task. In this case, the prevalence of trivial errors may be a useful on-line measure of comprehensibility.

Operationalizing Intelligibility and Comprehensibility: Response Latency

Munro and Derwing (1995) framed comprehensibility in terms of processing efficiency. Processing efficiency manifests in two ways: increased perceived effort, and response latency. Comprehensibility has been predominantly operationalized through the first method: scalar ratings made by listeners (Thomson, 2018). However, operationalizing comprehensibility as response latency provides a more direct measure of efficiency. However, only a handful of studies on foreign accented speech have taken response latency in account (Floccia et al., 2008; Hahn, 2004; Munro & Derwing, 1995; Wilson & Spaulding, 2010). On-line response latency is a key piece of the puzzle when it comes to determining processing efficiency. Replacing hand-written transcription with typed tasks provides a direct, on-line, simultaneous measure of both comprehensibility, in the form of keystroke latency, and intelligibility, in the form of response accuracy.

Research Question

This paper focuses on two typing-specific dependent variables: individual keystroke latency and corrected errors. The question it aims to answer is:

1. Do individual keystroke latencies and error correction rates differ for auditory stimuli presented in a foreign accent?

This analysis will help to answer the larger conceptual question of whether typing tasks provide a way of directly measuring comprehensibility and intelligibility simultaneously.

METHODS

Participants

In total, 116 university-aged English-speakers participated in this study. Of these, 64 were native speakers of English, 28 were native speakers of Mandarin, and 24 were from other language backgrounds. Participants were recruited at Brock University, McGill University, and University of Alberta (N=86), and via Amazon's Mechanical Turk crowdsourcing platform (N=30). All participants had normal or corrected to normal vision and reported using headphones while participating in the study. Participants were not screened on the basis of hearing-impairment. Non-Mandarin speaking participants reported having little to no familiarity with Chinese speakers of English.

Materials

The stimulus set used in this experiment was taken from a previous study examining phonological and morphological form overlap between Mandarin and English. Thus, half of the stimuli contained formal similarities with Mandarin words, such as 'moonlight' (which shares compound structure with its Chinese translation equivalent) and 'cookie' (which is phonologically similar to its loanword translation equivalent). The remaining stimuli were matched on lexical characteristics and selected from the English Lexicon Project (Balota et al., 2004). In total, 112 words were included in the final stimulus list.

Audio recordings were produced by four late, unbalanced Mandarin-English bilinguals and two monolingual native speakers of English. All talkers were graduate students at Brock University. Target stimuli were recorded in the carrier phrase "Now I say, _____", using a Marantz Professional PMD-561 handheld solid-state recording device with a pop cover. Recording was done continuously in a WhisperRoom sound isolation booth to reduce any potential environmental noise. Target words were extracted manually from the recorded audio files using Praat (v6.0.40). Once target stimuli had been extracted into individual audio files, they were normalized in Audacity (v2.3.0). Normalized audio files were assessed for recording quality, pops, or critical pronunciation errors by 2 monolingual native English-speaking raters. Recordings by two Mandarin speakers were removed because over half of their recordings were judged as unintelligible. In total, 448 audio recordings were created; 112 target words, each produced by two Mandarin-English bilinguals and by two monolingual English-speaking Canadians.

A questionnaire was adapted from the 'The Language Contact Profile' (Freed et al., 2004). Demographic information pertaining to age, gender, educational achievement, and language history were collected. In addition, participants were asked to self-rate their typing ability and indicate where they completed the experiment and the level of atmospheric noise in that location.

Procedures

The typing task used in this experiment was developed in PsychoPy3 (Peirce et al., 2019), an experiment development software capable of creating experiments in HTML and JavaScript. Experiments created in PsychoPy3 can be hosted on a unique URL and accessed by anyone with

a computer and an internet connection via their web-browser. Unlike many other online data collection methods, PsychoPy3 allows for the collection of response time data with millisecond precision.

Participants accessed the experiment by navigating to its URL. The experiment began with a digital consent form. Consent was indicated via button press and recorded in the data output. After being given instructions on how to complete the tasks, participants completed four practice trials before beginning the 112 experimental trials. Having completed all trials, participants responded to a short 12-item demographic questionnaire.

Each trial consisted of a two-second fixation point followed by the presentation of an auditory stimulus. Each auditory stimulus was a recording of one word spoken by either a native English-speaker or Mandarin-English bilingual. The order of stimulus presentation was randomized, and the accent condition of each lexical stimulus was counter-balanced across trials. Participants were instructed to type the auditory stimulus using the keyboard as soon as they recognized it. Every keystroke made by participants was recorded along with its response latency relative to the stimulus onset. The latency between each keystroke was then calculated. Participants' typed responses appeared in the middle of the screen and corrections could be made using the BACKSPACE prior to submission. Once participants finished typing, their response was submitted using the RETURN key. Upon completing all experiment trials, response data were converted to an Excel file and stored on the Gitlab repository.

Data Analysis

Outlier stimuli and participants were identified using normalized mean response latencies. Stimuli and participants with mean inter-keystroke intervals (IKSI) greater than three standard deviations from the mean were removed. Trials containing typing onset latencies >3000ms or <300ms or IKSIs >1000ms or <50ms were removed as they were not taken to reflect automatic on-line processing. Trials containing typing errors (even those subsequently corrected) were not included in the analysis for the same reason.

RESULTS

Accuracy

Accuracy of responses was collated conservatively. Trials containing any keystrokes not corresponding directly to the target were marked as error trials. This included trials containing corrected errors as it is not clear how typing latencies before and after corrected errors reflect on-line typed word processing. Further analysis of these trials is presented below. Accuracy data were analyzed using mixed effects logistic regression, which allows for regression analysis of a categorical dependent variable on the basis of both fixed and random effects. The random effects included in the model were target word and participant. The fixed effects included were native language of the listener, accent of the speaker, whole-word frequency, and word length. An interaction between native language and accent was added as it was predicted that accent may influence participants groups differently. A summary of fixed effects is presented in Table 1. The estimated probability of typing accuracy by accent and native language is plotted in Figure 1.

The model indicates that the native English speaking participants were more likely to respond accurately than both bilingual groups. The bilinguals' response accuracy was similar regardless of language background. All participant groups were less likely to respond correctly to Mandarin-accented stimuli. However, the effect of accent differed across participant groups. Native English speaking participants showed the greatest reduction in predicted accuracy for accent stimuli. Between the two bilingual groups, Mandarin-English bilingual participants showed the smallest reduction in predicted accuracy.

Table 2
Fixed effects of the predictors in the mixed effects logistic regression model of typed response accuracy.

Factor Names and Levels	**Est.**	**Std. Error**	**z value Pr(>|z|)**	**z value Pr(>|z|)**	**Sig.**
(Intercept)	0.72	0.52	1.39	0.16	***
Native Language: Mandarin	-1.55	0.22	-6.95	<0.001	***
Native Language: Other	-1.49	0.21	-7.02	<0.001	***
Accent: Mandarin	-1.26	0.08	-16.62	<0.001	***
Word Frequency (COBUILD Corpus)	0.45	0.19	2.41	0.02	*
Word Length	0.14	0.06	2.33	0.02	*
Accent: Mandarin * Native Language: Mandarin	0.82	0.12	4.55	<0.001	***
Accent: Mandarin * Native Language: Other	0.57	0.12	6.67	<0.001	***

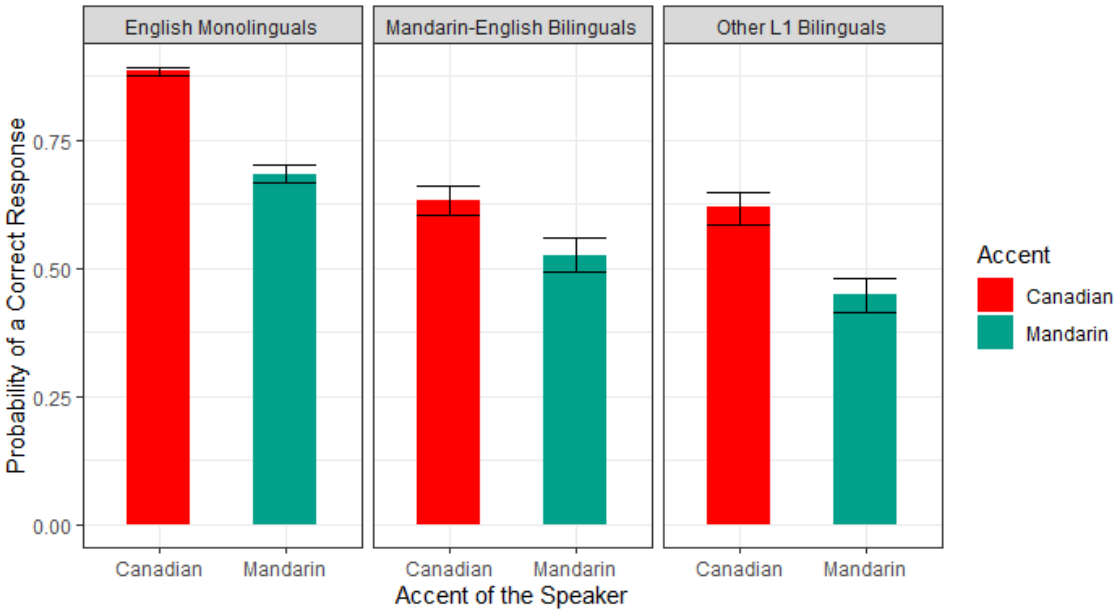

Figure 1. Log odds of correctly typing the target when produced by Mandarin-English bilingual speakers (Mandarin Accent) and Canadian monolingual English speakers (Canadian Accent) for participants grouped by native language background.

Inter-Keystroke Interval Latency

To determine whether foreign accent affected individual keystroke latency, typed responses were analyzed using a linear mixed effects model. The random effects included in the model were participant, target word, and letter typed. Letter typed was included to account for factors related to the relative positions of letter keys on a standard qwerty keyboard.

The main fixed effects in the model were accent of speaker (stimulus) and native language of the listener (participant). Control variables added to model were trigram frequency, backward and forward bigram frequency, word length, word frequency, keystroke order, and morpheme boundary. Trigram frequency refers to the frequency of three-letter combinations, consisting of the letter being typed, and the letter immediately before and after it. Bigram frequencies refers to the frequency of the two-letter combination formed from the current letter being typed and the preceding (backward) and proceeding (forward) letter. Higher frequency bigrams and trigrams are more automatized, and therefore, typed faster. Frequency of English bigrams and trigrams were taken from the Practical Cryptography website (Lyons, 2012). Keystroke position, target length, and morpheme boundary were included to account for variations due to internal lexical structure and typed production processes. A summary of fixed effects is presented in Table 2 and effect of accent is shown in Figure 2.

The estimates produced by the model indicate a main effect of accent on individual keystroke latencies. To determine whether accent differentially influenced keystroke latency at specific positions, an interaction between typing position and accent was added to the model. However, this interaction did not improve model performance and did not indicate any significant variation in keystroke latency across typing positions.

Table 3
Fixed effects of the predictors in the linear mixed effects regression model of inter-keystroke intervals.

Factor Names and Levels	**Estimate**	**Std. Error**	**df**	***t*-value**	***p***	**Sig.**
(Intercept)	5.36	0.08	389.9	70.63	<2e-16	***
Accent: Mandarin	0.02	0.01	21940	3.40	<0.001	***
Native Language: Mandarin	0.30	0.06	104.5	5.24	<0.001	***
Native Language: Other	0.31	0.06	106	5.10	<0.001	***
Trigram Frequency	-0.02	0.00	10880	-5.98	<0.001	***
Morpheme Boundary: Yes	0.37	0.01	18380	27.85	<0.001	***
Keystroke Position	-0.01	0.002	28790	-9.02	<0.001	***
Word Length	0.01	0.005	116.5	2.51	0.01	*

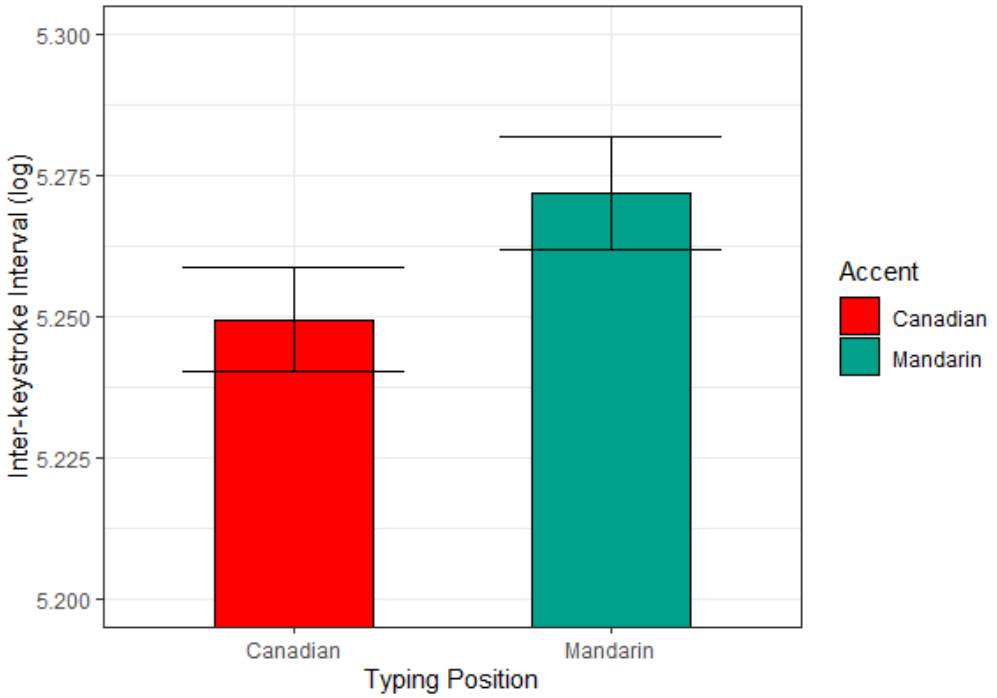

Figure 2. Individual keystroke latency for words spoken by Mandarin-English bilingual speakers (Mandarin Accent) and Canadian monolingual English speakers (Canadian Accent).

Corrected Errors

Corrected errors were analyzed as a binary variable at the trial level. Trials containing at least one typing error in which the correct target was eventually typed were labelled as corrected error trials. This did not take the number of corrected errors into account. The total number of trials containing corrected errors was analyzed as a proportion of the total number of trials containing errors. Analyzing errors in this manner provides a measure of the likelihood of recognizing on-line production errors.

Analysis of corrected errors was done using mixed effects logistic regression. The binary dependent variable was the corrected error status of each trial. Participants and target word were included in the model as random effects. The fixed effects included were native language of the listener and accent of the speaker. Control variables added to the model were whole-word frequency and word length. A summary of fixed effects is presented in Table 3 and the effects of accent by participant group is shown in Figure 3.

Results indicate that corrected errors were less likely to be made when the target was spoken in a foreign accent across all participant groups, though accent did show a differential effect across these groups. As can be seen in Figure 3, native English speakers were most likely to correct typing errors. They also showed the greatest difference across accent conditions. This effect was comparatively smaller for bilingual participants. Mandarin-English bilinguals were least likely to correct errors, although they did still show a difference across accent conditions.

Table 4
Fixed effects of the predictors in the mixed effects logistic regression model of corrected on-line typing errors.

Factor Names and Levels	Estimate	Std. Error	z value Pr(>|z|)	z value Pr(>|z|)	Sig.
(Intercept)	-1.92	0.48	-4.01	<0.001	***
Accent of Speaker: Chinese	-1.26	0.13	-9.58	<0.001	***
Native Language: Other	-0.86	0.30	-2.84	<0.001	**
Native Language: Chinese	-1.15	0.29	-3.93	<0.001	***
Word Frequency (COBUILD Corpus)	0.44	0.19	2.37	0.02	*
Word Length	0.26	0.06	4.26	<0.001	***
Accent: Mandarin * Native Language: Mandarin	0.81	0.20	4.15	<0.001	***
Accent: Mandarin * Native Language: Other	0.90	0.21	4.26	<0.001	***

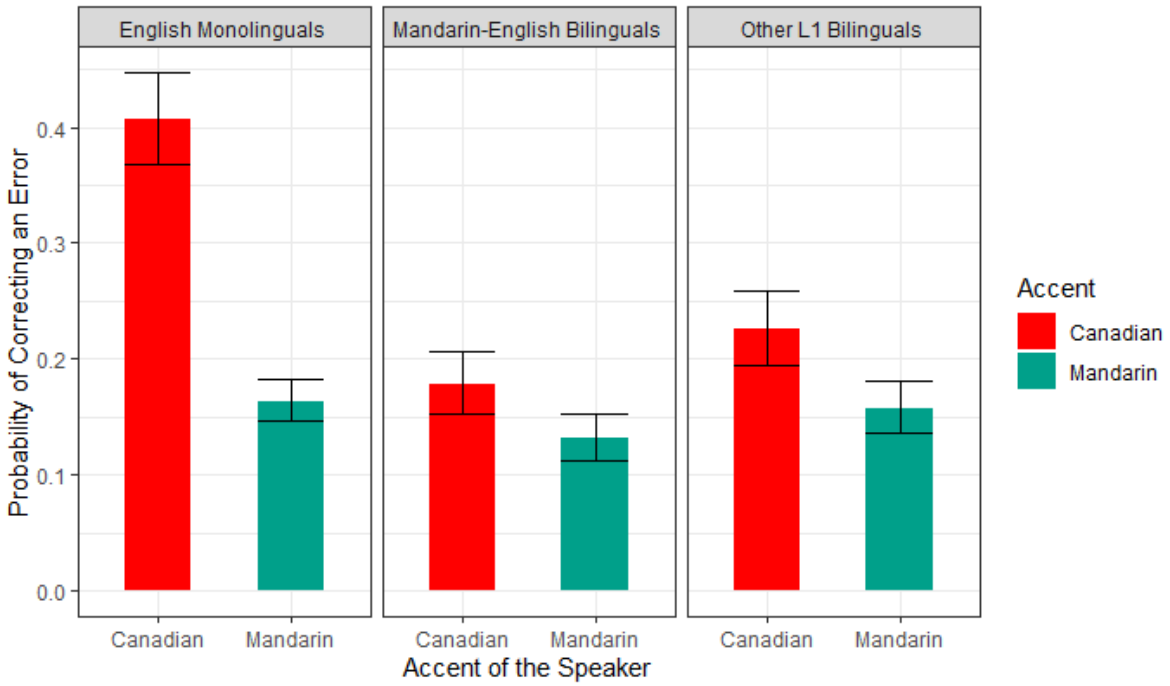

Figure 3. Log odds of correcting a typed production error during the transcription of speech produced by Mandarin-English bilingual speakers (Mandarin Accent) and Canadian monolingual English speakers (Canadian Accent) for participants groups by native language background.

DISCUSSION

This research introduced the typed transcription task as a new method in the study of foreign accent intelligibility. An auditory stimulus typed transcription task was conducted comparing response accuracy and latency for individual spoken words produced by either monolingual English speakers or Mandarin-English bilinguals. Corrected error rate and individual keystroke latency were analyzed to highlight the unique contribution of the typed transcription task over and above the traditional pen-and-paper method. Current technology allows the capture of typing latency,

which can be used as an analogue for comprehensibility. Results indicated significant differences in corrected error rates and keystroke latencies across all participant groups and between accent conditions.

Overall response accuracy patterns were consistent with previously reported results (Munro et al., 2006), indicating that foreign accented speech was less likely to be intelligible for all listeners, even those sharing L1 background with the speaker.

Keystroke Latency

Individual words spoken in a Chinese accent were typed slower, despite being recognized and typed correctly. This was indicated by significant increases in IKSI latency for Chinese-accented speech. This is consistent with previous studies which report greater response latencies to foreign accented targets in word recognition tasks (Munro & Derwing, 1995; Wilson & Spaulding, 2010). Longer latencies reflect the complexities involved in resolving phonological variation and mapping the incoming acoustic signal onto existing lexical representations. However, these results are unusual because by the time that the typed production of the word is initiated, the word should have been consciously recognized. The question is why the subsequent typed production of a word would be slower despite being successfully recognized.

The answer may require a consideration of the relative activation levels of target representations. During word recognition, multidimensional networks corresponding to the input become activated. For speech to be intelligible, the representation must reach a threshold activation level (Marslen-Wilson, 1987). However, the extent to which this threshold is exceeded may influence subsequent typed production. Representations contain an orthographic code used in the articulatory motor production of a word. Since foreign accented speech does not map as precisely onto the listener's phonological representation of the word, the relative activation of the network representation is relatively lower than for nonaccented speech. When the entire network representation of a word is relatively less activated, as in the case of foreign accented speech, it takes longer to access and execute the motor plan for that word. This may explain the greater IKSI latencies observed for foreign-accented speech. Comprehensibility may correspond not only to the cognitive processes involved in resolving unexpected phonological variation, but also the additional resources required to initiate processing after speech is recognized.

Corrected Errors

Typed errors are less likely to be corrected when transcribing foreign-accented speech. Regardless of the theoretical interpretation of this finding, a strong methodological consideration is highlighted here. Response accuracy typically operationalizes intelligibility only. However, this ignores the possibility that production errors, while trivial in terms of intelligibility, may be related to processing difficulty, and therefore, comprehensibility. According to the results reported here, participants are less likely to notice and correct errors when typing foreign accented speech. This suggests that trivial errors may not be as trivial as originally assumed. This finding highlights the importance of utilizing measures capable of capturing intelligibility and comprehensibility simultaneously.

CONCLUSION

This paper reported results of a transcription typing task aimed at simultaneously measuring intelligibility and comprehensibility of foreign speech. Both corrected errors and IKSI latency showed significant differences across accent conditions. Results support the conclusion that typed transcription tasks provide a viable method of directly measuring the intelligibility and comprehensibility of individual words on-line, as well as unique insights in the processes underlying foreign-accent word processing.

LIMITATIONS AND FUTURE DIRECTIONS

There are several limitations to this study. First, it was conducted online using recently released software. Thus, replication of these results using locally-run measurement software would strengthen the veracity of the results reported here. Second, foreign-accent was treated as a categorical variable in this study. Therefore, differences in typed responses to each speaker could only be attributed to accent, rather than some particular phonological quality of the speech produced. Lastly, whether the results reported here extend to the processing of foreign-accented speech beyond the individual word level will have to be determined by future studies.

Future studies into the typing of foreign accented speech can proceed in numerous directions. It is possible that typed production may be sensitive to phonological variation at certain typing positions within the word. More precise manipulation of phonological variation and phonetic analysis of auditory stimuli would help better understand the typed production of accented speech.

Further analysis of corrected errors could also shed light on the role of attentional resources involved in typing foreign accented speech. This study did not consider the position within the word where corrected errors occurred. It is possible that other word-internal structures and segment-specific phonological variations play a role in determining whether an error is noticed and corrected. Most importantly, utilizing typing data for analysis of sentence level transcription provides a wealth of possibilities for understanding the processing of accented speech.

Overall, this study provides a step in a slightly different and intriguing direction for research on the intelligibility and comprehensibility of foreign accented speech. The results reported here suggest that typed transcription tasks provide a new and exciting method of understanding the interplay between intelligibility and comprehensibility.

ABOUT THE AUTHOR

Jordan Gallant is a researcher currently based out of Brock University in Canada. His main focus is methodological innovation in psycholinguistic and second language research. His areas of interest center around the mental lexicon, lexical access, and typed word production.

REFERENCES

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology, 133,* 283–316.

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America, 114,* 1600–1610.

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19,* 1–16.

Floccia, C., Bulter, J., Goslin, J., & Ellis, L. (2008). Regional and foreign accent processing in English: Can listeners adapt? *J Psycholinguist Res, 38,* 379–412.

Freed, B. F., Dewey, D. P., & Segalowitz, N. (2004). The language contact profile. *Studies in Second Language Acquisition, 26(2),* 349–356.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38,* 201-223.

Libben, G., & Weber, S. (2014) Semantic transparency, compounding, and the nature of independent variables. In F. Rainer, W. Dressler, F. Gardani & H. C. Luschutzky (Eds.), *Morphology and meaning* (pp. 205–221). Amsterdam/Philadelphia: John Benjamins.

Lyons, J. (2012). *Practical cryptography*. [online] Practicalcryptography.com. Available at: http://practicalcryptography.com/ [Accessed 10 Nov. 2018].

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. *Cognition, 25,* 71-102.

Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38,* 289–306.

Munro, M. J., & Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition, 28*(1), 111.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51(1),* 195–203.

Thomson, R. (2018). Measurement of accentedness, intelligibility, and comprehensibility. In O. Kang, & A. Ginther, (Eds), *Assessment in Second Language Pronunciation* (pp. 11-29). New York, NY: Routledge.

Wilson, E. O., & Spaulding, T. J. (2010). Effects of noise and speech intelligibility on listener comprehension and processing time of Korean-accented English. *Journal of Speech, Language, and Hearing Research, 53,* 1543–1554.