Original research

Google translate in healthcare: preliminary evaluation of transcription, translation and speech synthesis accuracy

Jack Birkenbeuel ,[1] Helen Joyce,[1] Ronald Sahyouni,[2] Dillon Cheung,[1] Marlon Maducdoc,[1] Navid Mostaghni,[1] Sammy Sahyouni,[1] Hamid Djalilian,[1] Jefferson Chen,[3] Harrison Lin[1]

► Supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjinnov-2019-000347).

[1]Otolaryngology - Head and Neck Surgery, University of California Irvine, Irvine, California, USA
[2]Neurosurgery, University of California San Diego, La Jolla, California, USA
[3]Neurosurgery, University of California Irvine, Irvine, California, USA

Correspondence to
Dr Harrison Lin, Otolaryngology - Head and Neck Surgery, University of California Irvine, Irvine, CA 92697, USA; harrison.lin@uci.edu

JB, HJ and RS contributed equally.

Received 14 February 2019
Revised 25 January 2021
Accepted 27 January 2021
Published Online First 2 March 2021

To cite: Birkenbeuel J, Joyce H, Sahyouni R, *et al*. *BMJ Innov* 2021;**7**:422–429.

ABSTRACT

Objective To assess the ability of Google Translate (GT) to accurately interpret single sentences and series of sentences commonly used in healthcare encounters from English to Spanish.

Design English-speaking volunteers used GT to interpret a list of 83 commonly used sentences and series of sentences of different lengths containing both medical and non-medical terminology. A certified medical interpreter evaluated whether the meaning of these sentences was preserved.

Participants Eighteen English-speaking subjects (nine males and nine females), with a mean age of 36 years, volunteered for this study to read sentences.

Main outcome measures The accuracy of GTs (1) real-time voice recognition (ie, transcription) of English sentences, (2) real-time translation of these transcribed English sentences to Spanish, and (3) GTs speech synthesis ability to preserve the meaning of spoken English sentences after translation to Spanish.

Results Speech synthesis accuracy, with preservation of the original English-spoken sentence(s), was 89.4% for single sentences with ≤8 words; 90.6% for single sentences with >8 words; 52.2% for two sentences and 26.6% for three sentences. Furthermore, the number of transcription and translation errors per sentence(s) significantly increased with the number of sentences ($p<0.05$).

Conclusions Despite the fact that GTs accuracy was widely variable and dependent on the length of the spoken sentence(s), GT is readily accessible, has no associated monetary costs, and offers nearly immediate interpretation services. As such, it has the potential to routinely facilitate effective one-way oral communication between English-speaking physicians and Spanish-speaking patients with limited English proficiency.

Summary box

What are the new findings?

- This is the first study to evaluate Google Translate's ability to interpret medical phrases from English to Spanish.
- Google Translate's ability to accurately interpret medically related sentences from English to Spanish depends on the length of the sentence(s) spoken into the application, ranging from 90% for single sentences to 27% for three sentences.
- Google Translate's interpretation accuracy significantly decreases with the use of one or more medical terms.

How might it impact on healthcare in the future?

- In under-resourced settings or settings where in-person interpreters are unavailable, our study suggests that Google Translate has appropriate interpretation rates to allow for effective one way, English to Spanish communication with limited English proficiency patients when using single sentences.

INTRODUCTION

Census data from 2013 revealed that ~21% of the US patient population (61.6 million individuals) spoke a language other than English at home, with 41% of these patients (~8.5% of the total US population) having limited English proficiency (LEP).[1] In California, 58%–66% of Asian-Americans and 40% of Hispanics have LEP, necessitating certified medical interpreters (CMI) or remote interpreters (ie, both phone and video interpreters) to mediate language discordance between patients and healthcare professionals.[2 3] Constraints to CMIs and remote interpreters include the associated healthcare costs and low availability compared with demand. Although in-person interpretation remains the gold standard, recent technological advancements in online translation services and near real-time translation and interpretation technologies warrant a thorough evaluation of their roles in medical communication with LEP patients.

A limited number of reports have evaluated the accuracy of online translation (written communication) and interpretation (oral communication) services in medical settings. Kaliyadan and Gopinathan Pillai found that patients using only online translation and interpretation services were equally satisfied with their doctor–patient interaction when compared with those who used a CMI.[4] In 2014, Patil and Davies assessed the ability of Google Translate (GT), a packaged combination of online translation services and near real-time translation and interpretation, to translate ten commonly used medical statements in 26 different languages.[5] They concluded that online translation and interpretation services are useful supplements to CMI when these resources are not available. Since the publication of these studies, advancements in machine learning and the use of neural machine translation systems (NMT) have significantly enhanced translation and interpretation capabilities and error reduction in these online services.[6] In 2016, GT transitioned to using an NMT, which improves translation and interpretation accuracy by making word choices based on pattern discovery and context interpretation.[7 8] Since the algorithm was updated, studies have reported GT as an effective tool for both translating text-written discharge instructions from English to Spanish and Chinese, and text-written portal messages to the physicians of Portuguese-speaking patients.[9 10]

However, at present, the efficacy of GTs ability to transcribe, translate and synthesise speech from English spoken sentences into Spanish spoken sentences, has not been thoroughly evaluated in a medical setting following NMT implementation, as prior studies only evaluated written messages. Together the processes of transcription, translation and speech synthesis are analogous to the human process of interpretation, which is the conversion of oral communication from one language to another. Translation is the process of converting written language from one language to another. The potential benefits of online translation and interpretation services and near real-time technology include widespread accessibility, negligible or minimal infrastructure costs, and nearly instantaneous augmentation of interlinguistic patient–physician communication. Based on these potential benefits, we evaluated GTs capacity to transcribe, translate and synthesise speech from spoken English to spoken Spanish in a healthcare setting.

METHODS

Sentences

With institutional review board exemption, 83 sentences and series of sentences commonly used during patient–physician interactions for this study were chosen by author consensus (online supplemental appendix 1). The 83 sentence(s) were considered of adequate power, as an a priori power analysis performed with a power level of 0.80, effect size of 0.50 and significance level of 0.05 yielded a minimum sample size of 64. Four researchers (JB, HJ, RS, and HL) classified each sentence(s) as having no medical terms or one or more. After consensus, the four researchers defined medical terms as any word used by medical professionals to describe anatomy, physiology, medical treatments, procedures, diseases and pharmacology that are not commonly used or understood by patients without a medical background. We did not classify known parts of the human body, like ears, as medical terms. We classified sentences as having (1) one sentence and eight words or fewer, (2) one sentence and more than eight words, (3) series of two sentences and (4) series of three sentences.

Experimental design

The experiment occurred from June to July 2018. We standardised testing conditions for all participants. We downloaded the GT application from the App Store onto an iPad running iOS 12.1.1 to eliminate variation in device hardware and carrier reception. High-throughput Wi-Fi was used during the experiments to ensure no transcription or translation delay. Data were collected in a quiet room, and iPads were secured to prevent movement. Participants were instructed to sit two feet away from the iPad. Before the trial, participants were given a printed copy of the sentences to review to mitigate unintentional reading mistakes or improper enunciation of words (online supplemental appendix 1). A researcher supervised the subjects during the trial. Participants were asked to speak with normal cadence. GT provided text transcription for each spoken sentence(s) (online supplemental appendix 2), the translation of the transcribed English sentences into Spanish in text form (online supplemental appendix 2), and the translated Spanish text into speech form, denoted as 'speech synthesis,' which is the artificial production of human speech. The entire trial was screen recorded with the microphone enabled. This process was repeated for all 83 sentence(s).

We separated the collected data in the following components: (1) voice and text recognition (transcription step) of each initial English sentence(s); (2) Spanish translation of each transcribed sentence(s) (translation step); (3) speech synthesis of each translated Spanish text sentence(s) spoken through GT (speech synthesis step). See online supplemental appendix 2 for design layout. All three steps, when combined, allowed the researchers to assess GTs ability to interpret English to Spanish.

Transcription, translation and speech synthesis analysis occurred as follows: for all four sentence types, the mean accuracy by sentence type was generated for all 18 participants (ie, sentence lengths). For example, the transcription accuracy for all single sentences <8 words (n=44) were averaged for each participant. This was done separately for translation and speech synthesis steps. We then averaged the mean number of errors for each sentence type for all 18 particpants together, providing the overall transcription, translation and speech synthesis accuracy.

Transcription analysis

Analysis was completed in July 2018. The sentence(s) in online supplemental appendix 1 were compared with each transcribed sentence(s) to assess the accuracy of speech to text recognition. Inappropriate word insertions, deletions and substitutions were counted as errors for both transcription and translation. For example, if the expression 'side effects' was deleted, two errors were counted. Examples of transcription errors are in online supplemental appendix 3. Transcribed sentences in GT do not include punctuation; hence, punctuation was not included in the accuracy assessment. Two team members (JB and RS) independently annotated the errors. Interobserver disagreement was resolved with discussion with HJ. For each of the 83 sentences or series of sentences, the resulting transcriptions of all participants were averaged to provide a total number of errors for each sentence and an overall transcription accuracy rate. Outcome variables included overall transcription accuracy rate (%) for the entire sentence list, transcription accuracy between sentences with no medical terms or one or more, and transcription accuracies (%) and mean number of errors specified for each sentence type (ie, single sentence ≤8 words, single sentence >8 words, two sentences and three sentences) to determine if transcription accuracy differed based on sentence length.

Translation analysis

After the first step, we proceeded with the translation step. It is important to note that translation is classically defined as the conversion of written communication between languages, which is why the researchers called this step the translation analysis. As such, English sentence(s) were translated to Spanish through GT. Subsequently, the same CMI translated these Spanish sentences back to English. The back-translation of each translated sentence was then compared with the original English sentence. The same two researchers independently reviewed each translation with a third team member present to resolve any discrepancies. Sentences were reverse-translated back to English based on the methodologies used in existing studies in the literature.[4 5] Because GT only includes punctuation in translated sentences but not in transcribed sentences, punctuation mistakes were excluded from the analysis. Outcome variables included overall translation accuracy rate (%) for the entire sentence list, translation accuracy rate between sentences with no medical terms or one or more, and translation accuracies (%) and mean number of errors categorised by sentence type (ie, single sentence ≤8 words, single sentence >8 words, two sentences and three sentences) to determine if translation accuracy differed based on sentence length.

Speech synthesis rate analysis

The same CMI assessed GTs ability to preserve the meaning of each original sentence(s) through speech synthesis, the artificial production of human speech. The CMI assessed GTs ability to read the translated text in a way that preserved its meaning. The CMI listened to the recorded trial and annotated GTs spoken sentence(s). This sentence(s) was then converted back into English by the same CMI. This back-translation was separate from the back-translation used in the translation analysis step. Finally, the CMI compared each GT speech synthesis output to its initial English counterpart from the sentence list. The CMI then evaluated the resulting interpretation and evaluated meaning preservation of each sentence(s) in each trial as a binary 'yes' or 'no'. The CMI determined whether or not the meaning of a sentence(s) was preserved based on the following criteria: (1) whether GT allowed the speaker to finish their intended sentence(s) (interruption errors); (2) whether GT inflected the sentence to correctly convey the intention of the sentence(s) (eg, GT inflected a statement if the speaker inflected a question, termed inflection errors); (3) whether GT correctly detected sentence breaks between consecutive sentences (sentence break errors); and (4) whether GT successfully audibly spoke the translated sentence(s) (failure to synthesise speech). In certain cases, GT stopped speaking before completing the entire translated sentence(s) and this was denoted as failure to synthesise speech error.

Statistical analysis

All statistical analyses were performed using PASW 18.0 (SPSS, Chicago, Illinois, USA). Data were normally distributed, so we compared means of two continuous variables using independent sample t-tests, one-way analysis of variance (ANOVA) (with Tukey's post hoc analysis) to compare means of more than two continuous variables, and χ^2 tests for categorical

Table 1 Descriptive statistics table of the accuracy of transcription, translation and meaning preservation overall by sentence length

Outcomes	Overall rate (%)	Error rate among sentences
Transcription accuracy	98.3 (range: 80.24–100) SD: 2.8	Sentence with ≤8 words: 99.2% Sentence with >8 words: 98.7% Two sentences: 96.8% Three sentences: 94.2%
Translation accuracy	92.2 (range: 60–100) SD: 10.7	Sentence with ≤8 words: 93.6% Sentence with >8 words: 97.8% Two sentences: 81.0% Three sentences: 83.9%
Interpretation rate after speech synthesis	77.6% (proportion of times speech synthesis preserved meaning of translated text)	Sentence with ≤8 words: 89.4% Sentence with >8 words: 90.6% Two sentences: 52.2% Three sentences: 26.6%

variables. A p value of 0.05 or less was considered statistically significant.

RESULTS

Demographics

Eighteen individuals (nine males and nine females) participated in our testing of GT. Mean age was 36 years (range, 20–74 years; SD 18). Ethnicities of 18 participants included Caucasian (n=14), Asian or Pacific Islander (n=2) and other (n=2). All participants spoke English as a first language and American English from the western region of the USA. This is important given that phonological variations between regions can affect GTs speech recognition.[11 12]

Transcription accuracy

GT was 98.3% accurate across all sentence(s) (range: 80.24%–100%; SD 2.8), with no significant difference between males and females (p=0.52) (table 1). GT transcription accuracies for single sentences ≤8 words, single sentences>8 words, two sentences and three sentences were 99.2%, 98.7%, 96.8% and 94.2%, respectively (table 1).

ANOVA found a statistically significant difference between groups (p<0.001), with a statistically significant difference between single sentence ≤8 words and both two sentences (p<0.001) and three sentences (p<0.001) and a statistically significant difference between single sentences >8 words and three sentences (p<0.001) on post hoc analysis. All other comparisons did not meet statistical significance (table 2). Table 3 lists the mean number of transcription errors per sentence for each sentence length. We found a stepwise increase in the mean number of transcription errors with increased number of sentences (all p<0.05) (table 4). When comparing sentences with no or one or more medical term(s), the presence of one or more medical term(s) significantly decreased transcription accuracy (p=0.033) (table 5).

Translation accuracy

Overall translation accuracy of GT was 92.2% (range: 60%–100%; SD 6.7), with no significant difference between males and females (p=0.08) (table 1). Translation accuracy for single sentences ≤8 words, single

Table 2 Tukey's post hoc transcription analysis and translation analysis

		Transcription analysis			Translation analysis		
			95% CI			95% CI	
Input sentence	Comparative sentence	P value	Lower bound	Upper bound	P value	Lower bound	Upper bound
Sentence ≤8 words	Sentence >8 words	0.836	−9.84	0.86	0.404	−10.68	2.70
	2 sentences	0.023*	1.54	24.16	0.002*	3.92	21.78
	3 sentences	<0.001*	4.09	15.92	0.015*	1.45	18.6
Sentence >8 words	Sentence ≤8 words	0.836	−0.86	8.84	0.404	−2.70	10.7
	2 sentences	0.173	5.84	27.84	<0.001*	6.94	26.7
	3 sentences	<0.001*	9.40	18.58	0.001*	4.44	23.54
2 sentences	Sentence ≤8 words	0.023*	−0.86	8.84	0.002*	−21.78	−3.92
	Sentence >8 words	0.173	5.84	27.84	<0.001*	−27.73	−6.95
	3 sentences	0.063	9.40	18.58	0.910	−14.09	8.39
3 sentences	Sentence ≤8 words	0.000*	−24.16	−1.54	0.015*	−18.55	−1.45
	Sentence >8 words	0.000*	−27.84	−5.84	0.001*	−23.55	−4.44
	2 sentences	0.063	−14.06	8.36	0.910	−8.39	14.09

*P values < 0.05 were considered statistically significant.

Table 3 Description of the mean number of transcription and translation errors by sentence length

	Mean no of transcription errors per sentence(s)	SD	95% CI		Mean no of translation errors per sentence(s)	SD	95% CI	
			Lower bound	Upper bound			Lower bound	Upper bound
One sentence (<8 words)	0.047	0.118	0.0389	0.0551	1.59	0.087	1.584	1.596
One sentence (>8 words)	0.173	0.312	0.1398	0.2062	1.37	0.603	1.3059	1.4341
Two sentences	0.698	1.16	0.518	0.878	5.05	2.089	4.7258	5.3742
Three sentences	2.70	3.15	2.2368	3.1632	7.68	2.136	7.3659	7.9941

sentences >8 words, two sentences and three sentences were 93.9%, 97.8%, 81.0% and 83.9%, respectively (table 1). These four different sentence lengths were compared through ANOVA and Tukey's post hoc analysis. ANOVA demonstrated a significant difference between groups ($p<0.001$). Post hoc Tukey's analysis indicated a significant difference between single sentences ≤8 words and two sentences ($p=0.002$) and three sentences ($p=0.015$). The analysis demonstrated a significant difference between single sentences >8 words and two sentences ($p<0.001$) and three sentences ($p=0.001$). All other between-group comparisons were nonsignificant (table 2). Table 3 lists the mean number of translation errors per sentence for each sentence length. We report a stepwise increase in the mean number of translation errors between single sentences and two and three sentences (all $p<0.05$), with no significant differences between two and three sentences (table 4). However, there was no significant decrease in translation accuracy between sentences with no or one or more medical term(s) ($p=0.293$) (table 5).

Speech synthesis accuracy

GTs speech synthesis accurately interpreted sentence(s) 77.6% of the time. GTs speech synthesis accurately interpreted single sentences ≤8 words 89.4% of the time; single sentences >8 words 90.6% of the time; two consecutive sentences 52.2% of the time; and three consecutive sentences 26.6% of the time (online supplemental appendix 4). χ^2 tests found a statistically significant difference among all four sentences of different lengths ($p<0.001$). When assessing the influence of medical terms on speech synthesis accuracy, the presence of one or more medical term (mean: 72.9%) significantly decreased the speech synthesis accuracy compared with sentences with no medical terms (mean: 81.9%) ($p<0.001$). The variability in error rates categorised by sentence length are in table 6. The per cent of speech synthesis error types leading to loss of meaning preservation is shown in online supplemental appendix 5.

Table 4 Comparison of the mean number of transcription and translation errors between sentence lengths

Comparators		Transcription	Translation
		P value	P value
One sentence (<8 words)	One sentence (>8 words)	0.130	0.139
One sentence (<8 words)	Two sentences	0.028*	<0.001*
One sentence (<8 words)	Three sentences	0.001*	<0.001*
One sentence (>8 words)	Two sentences	0.081	<0.001*
One sentence (>8 words)	Three sentences	0.002*	<0.001*
Two sentences	Three sentences	0.019*	0.075

*P values < 0.05 were considered statistically significant.

DISCUSSION

This study demonstrates that GT is a useful tool for one-way medical interpretation of single sentences, and although GT accuracy is limited by sentence length, its accessibility, reliability, accuracy and affordability warrant further investigation into its utility to facilitate two-way communication for patients with LEP. We found that GTs ability to accurately interpret medically related sentences from English to Spanish depends on the length of the sentence(s) spoken into the application. As the number of sentences increased, interpretation accuracy decreased in a stepwise manner ($p<0.001$). Similarly, the mean number of transcription and translation errors increased stepwise with more sentences ($p<0.05$). We found the most substantial decline in accuracy (between transcription, translation and speech synthesis) to be between the translation and speech synthesis step, with overall accuracy dropping 14.6%, compared with a 6.1% decrease from transcription to translation. An improvement in GTs ability to accurately synthesise speech from the translated sentence(s) could increase its accuracy and potential utility as a form of one-way interpretation from English to Spanish in a medical setting.

In 2016, GT transitioned from a phrase-based statistical machine translation and interpretation system to an NMT system, showing an unprecendented improvement in BLEU score (bilingual evaluation understudy).[7,8] Prior to the update, a 2014 study evaluated GTs accuracy in translating 10 written English medical statements into 26 languages, finding that 57.7% of the translations were correct.[5] A 2013 study tested GTs ability to translate German text from a neonatal intensive care unit brochure to English, Portuguese, and

Table 5 Assessment of the influence of medical terms on transcription, translation and speech synthesis accuracy

					95% CI		
Factors	**Presence of medical terms**	**N**	**Mean % Correct**	**SD**	**Lower bound**	**Upper bound**	**P value**
Txn	None	954 (53*18 participants)	98.87	1.35	98.78	98.96	0.033*
	1+	540 (30*18 participants)	97.18	4.05	96.84	97.52	
Tln	None	954 (53*18 participants)	93.10	10.39	92.44	93.76	0.293
	1+	540 (30*18 participants)	90.51	11.25	89.56	91.46	
Speech Synthesis	None	954 (53*18 participants)	81.9	0.587	79.46	84.34	<0.001*
	1+	540 (30*18 participants)	72.9	0.449	69.15	76.65	

*P values < 0.05 were considered statistically significant.
Tln, translation; Txn, transcription.

Arabic and found that an average of 42% of sentences was translated correctly.[13]

Since GTs update, other studies have examined the translation accuracy of GT with medical text, but did not assess GTs ability to interpret sentences of varying lengths and detect fluctuations inherent in spoken speech during real-time interpretation.[5 9 10] To our knowledge, this is the first interpretation study to assess its accuracy when medically related sentences are spoken into the app. Khoong *et al* reported accurate translation of written discharge instructions from English to both Spanish and Chinese.[9] Rodriguez *et al* investigated GTs ability to translate written Portuguese portal messages to English, reporting GTs translation as non-inferior to human translation on all but one question.[10] We report similar findings with GT transcribing English speech to Spanish text. Our study differs as the prior studies used written communication (translation) and our study used oral communication (interpretation). Furthermore, our study compares accuracy by sentence length. By doing so, we have demonstrated GTs stepwise decrease in accuracy and increase in mean number of errors with increased number of sentences. As we demonstrate no difference in accuracy between single sentences <or> 8 words, we recommend future studies combine these groups to evaluate single sentence accuracy as one group.

We also investigated the inclusion of medical terms on GT accuracy. We report a significant decrease in accuracy in both the transcription and speech synthesis steps with the use of one or more medical terms. Similarly, Khoong *et al* reported a reduction in accuracy in Chinese when medical terminology was used.[9] Here, we corroborate prior findings that the use of one or more medical terms reduces GTs transcription ability and novel findings demonstrating decreased interpretation accuracy with one or more medical term.

A 2018 article reported that many healthcare providers use GT because it is easy to use and because accessing in-person and even online interpretation services is difficult.[14] At present, GTs use in a medical setting is limited by its inability to self-correct errors, which can lead to misunderstandings with potentially grave medicolegal consequences. However, as GT continues to improve, and if it is used with single sentences spoken with clear inflection, GT has the potential to improve patient care in situations when language-concordant providers and in-person and remote interpreters are unavailable (eg, under-resourced settings).

We emphasise that we are not proposing that GT replace standard in-person interpretation services, particularly for purposes of consent or legal documentation. The general consensus is that in-person interpreters provide a superior service due to improved satisfaction and patients' understanding of their diagnoses.[15–17] At this point, GT does not integrate the nuances of interpersonal communication that are better detected and understood by trained, in-person interpreters. However, in under-resourced settings or settings when in-person interpreters are unavailable, our study suggests that GT has appropriate interpretation

Table 6 Variability in error rates broken down by sentence length

Error rates	Single sentence: eight words or fewer	Single sentence: nine words or more	Two sentences	Three sentences	P value
Speech synthesis: 1. Inflection errors	4.2%	0.0%	14.4%	4.0%	<0.001*
Speech synthesis 2. Interruption errors	0.0%	0.0%	10.6%	17.2%	<0.001*
Speech synthesis 3. Sentence break errors	0.0%	0.0%	5.0%	28.7%	<0.001*
Speech synthesis 4. Failure to synthesise speech error	0.0%	0.0%	0.0%	7.5%	<0.001*

*P values < 0.05 were considered statistically significant.

rates to allow for effective one way, English to Spanish communication with LEP patients. Based on finding, a stepwise decrease in accuracy and stepwise increase in the number of errors with increased number of sentences, we recommend that providers speak in single sentences, as the single sentence accuracy of 90% accuracy significantly drops with two or more sentences. Therefore, we do not recommend speaking in two or three consecutive sentences, as this could harm patients if information from providers is not accurately interpreted. While we used back-translation to assess transcription, translation and speech synthesis errors, we recognise this has the potential to introduce errors separate from GT errors. While prior studies have also used back-translation,[4 5] we recognise future studies may also assess accuracy with methods separate from back-translation. Future studies may also benefit from further investigation of GTs accuracy based on medical terminology and readability scores as the American Medical Association recommends patient reading education material to be no higher than a sixth-grade reading level (≤6).[18]

Our study has several limitations. First, there are inherent limitations to using a patient non-secure platform when aiding in medical care. If GT were to be used in a medical setting, providers and patients would need assurance on the HIPAA compliance of this application to ensure patient confidentiality. Second, this study only examined GTs accuracy from English to Spanish. Further studies are needed to assess GTs ability to accurately interpret Spanish to English and to withstand a patient-provider conversation in the conversation mode of the application. Despite this limitation, this study adds value as a primary proof of concept for GTs ability to help physicians provide one-way instructions to patients with LEP. This proof of concept can be strengthened in future studies. Third, this study used one CMI to assess meaning preservation, which has the potential to introduce error without interobserver agreement. However, we used a CMI with extensive interpretation training, which is a significant improvement to previous studies, which used clinicians without mention of medical interpretation certification. Finally, we did not evaluate GTs capacities across languages other than Spanish. Before testing the implementation of GT for patient care, the application would need to be tested in clinical settings to determine Spanish-to-English accuracy, patient and provider satisfaction, ease-of-use, and accuracy in a less controlled environment like that of a hospital room.

CONCLUSIONS

GT performs best with medical language from English to Spanish when single sentences are spoken. This study demonstrates that GT has the potential to improve access to language services for Spanish-speaking patients with LEP. Further studies should assess GTs ability to accurately interpret Spanish to English, thereby providing information on two-way communication. Future studies are needed to determine the extent of its capabilities in clinical settings.

Acknowledgements We thank the numerous references listed in this manuscript that guided our research.

Contributors RS conceived of the presented idea. JB and HJ conceived and planned the experiments. JB, HJ, DC, MM, NM and SS carried out the experiments. JB and HJ collected and analysed the data. JB and HJ wrote the manuscript with support from all authors. HD, JC and HL supervised the project.

Funding RS was funded, in part, by NIH grant number T32 GM008620-17.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available on reasonable request. Deidentified participant data are available from researcher JB, whose ORCID identifier is 0000-0002-2133-7708. Reuse is not permitted unless previously discussed with researcher.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iD
Jack Birkenbeuel http://orcid.org/0000-0002-2133-7708

REFERENCES

1 U.S. Census Bureau. 2013 American Community Survey. In: Ruggles S, Trent Alexander J, Genadek K, *et al*, eds. *Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]*. Minneapolis: University of Minnesota, 2010.

2 S-P T, Yip M-P, Chun A. Development of intervention materials for individuals with limited English proficiency: lessons learned from "Colorectal Cancer Screening in Chinese Americans". *Med Care* 2008;46.

3 Shi L, Lebrun LA, Tsai J. The influence of English proficiency on access to care. *Ethn Health* 2009;14:625–42.

4 Kaliyadan F, Gopinathan Pillai S. The use of Google language tools as an interpretation aid in cross-cultural doctor-patient interaction: a pilot study. *Inform Prim Care* 2010;18:141–3.

5 Patil S, Davies P. Use of Google translate in medical communication: evaluation of accuracy. *BMJ* 2014;349:g7392.

6 Castelvecchi D. Deep learning boosts Google translate tool. *Nature* 2016.

7 Wu Y, Schuster M, Chen Z. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*, 2016.

8 Johnson M, Schuster M, QV L. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *arXiv* 2016.

9 Khoong EC, Steinbrook E, Brown C, *et al*. Assessing the use of Google translate for Spanish and Chinese translations of emergency department discharge instructions. *JAMA Intern Med* 2019;179:580–2.

10 Rodriguez JA, Fossa A, Mishuris R, *et al*. Bridging the language gap in patient portals: an evaluation of Google translate. *J Gen Intern Med* 2020. doi:10.1007/s11606-020-05719-z. [Epub ahead of print: 19 Feb 2020].

11 Labov W, Ash S, Boberg C. *The atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter, 2008.

12 Carver CM. *American regional dialects: a word geography*. University of Michigan Press, 1989.

13 Börner N, Sponholz S, König K, *et al*. [Google translate is not sufficient to overcome language barriers in neonatal medicine]. *Klin Padiatr* 2013;225:413–7.

14 Moberly T. Doctors choose Google translate to communicate with patients because of easy access. *BMJ* 2018;362:k3974.

15 Karliner LS, Jacobs EA, Chen AH, *et al*. Do professional interpreters improve clinical care for patients with limited English proficiency? A systematic review of the literature. *Health Serv Res* 2007;42:727–54.

16 Flores G. The impact of medical interpreter services on the quality of health care: a systematic review. *Med Care Res Rev* 2005;62:255–99.

17 Garcia EA, Roy LC, Okada PJ, *et al*. A comparison of the influence of Hospital-Trained, AD hoc, and telephone interpreters on perceived satisfaction of limited English-Proficient parents presenting to a pediatric emergency department. *Pediatr Emerg Care* 2004;20:373–8.

18 Weiss BD. Health literacy. *Am Med Assoc* 2003;253.