BabelDr vs Google Translate: a user study at Geneva University Hospitals (HUG)

Pierrette Bouillon[1], **Johanna Gerlach**[1], **Hervé Spechbach**[2], **Nikos Tsourakis**[1], **Sonia Halimi**[1]
[1] FTI/TIM, University of Geneva
{pierrette.bouillon, johanna.gerlach, nikolaos.tsourakis, sonia.halimi}@unige.ch
[2] Geneva University Hospital (HUG)
herve.spechbach@hcuge.ch

Abstract

This paper presents a user study carried out at Geneva University Hospitals (HUG) where we compared BabelDr, a flexible phraselator, with Google Translate (GT). French speaking doctors were asked to use both systems to diagnose Arabic speaking patients. We report on the user's interactions with both systems, the quality of translation, the participant's ability to reach a diagnosis with the two systems as well as user satisfaction.

1 Introduction

In the context of the current European refugee crisis, hospitals are more and more often obliged to deal with patients who have no language in common with the staff, and may also fail to share the same culture. For example, at the Geneva University Hospitals (HUG), Geneva's main hospital, 52% of the patients are foreigners and 10% speak no French at all. In 2015, the languages which caused most problems were Tigrinya, Arabic and Farsi; refugees from Eritrea, Syria and Afghanistan make up about 60% of all new demands for asylum in the area (SEM Newsletter, October 2015). The problems are not only linguistic. Cultural differences mean that these patients may have different conceptualizations of medicine, health care (Hacker et al., 2015), illness and treatment (Priebe et al., 2011). A situation of this kind, with barriers in language, culture and medical understanding, creates serious problems for quality, security and equitability of medical care, as has been pointed out by several researchers ((Flores et al., 2003) and (Wasseman et al., 2014)). Others underline the negative impact these issues have on health care costs (Jacobs et al., 2007).

In absence of qualified interpreters, a number of solutions are available today, but they all have their drawbacks. Phone-based interpreter services are very expensive (3 CHF/minute in Switzerland), not always available for some languages, and known to be less satisfactory than face-to-face interaction through a physically present interpreter (Wu et al., 2014). Google Translate (GT), increasingly often used when no other alternatives exist, is known to be unreliable for medical communication (Patil and Davies, 2014). Other tools like MediBabble and Universal Doctor have been developed specifically for the medical diagnosis scenario and translate a set of fixed questions, but are technically unsophisticated and content cannot easily be changed. Similar remarks apply to medical resources developed for refugees in conflict zones, like the *Medical Handbook for Refugees*[1], which are non-interactive databases.

BabelDr[2] is a joint project of Geneva University's Faculty of Translation and Interpreting (FTI) and Geneva University Hospitals (HUG) which specifically addresses this problem of lack of qualified interpreters in hospitals in languages spoken by refugees. The BabelDr application can be characterised as a flexible speech-enabled phrasebook (Rayner et al., 2016). Semantic coverage consists of a prespecified set of utterance-types, but users can use a wide variety of surface forms when speaking to the system. Each utterance-type is associated with a canonical source language version, which is rendered into the target languages

[1] https://www.refugeephrasebook.de/medical%20phrases/
[2] http://babeldr.unige.ch/

by suitably qualified translation experts. The central design goals are to ensure that a) translations are reliable, b) new target languages can easily be added, enabling flexibility in the face of changing patient demographics and c) content can be changed depending on the context.

In this paper, we present a user study done at Geneva University Hospitals (HUG) where we compared the baseline version of BabelDr with the online desktop version of GT in real hospital settings. We report on the 1) interactions in both systems, 2) quality of translation and 3) impact on diagnosis and satisfaction. Our hypotheses are that GT is not precise enough for this domain and that BabelDr is robust enough to make the diagnosis possible. Section 2 presents BabelDr in more detail; Section 3 describes the experiment and Section 4 the results.

2 BabelDr

The baseline version of BabelDr used for this experiment has been designed to assist in triaging of non-French-speaking patients visiting HUG's A&E department. It allows medical professionals to perform a preliminary medical examination dialogue, using a decision-tree method, to determine the nature of the patient's problem and the appropriate action to take. The coverage of the current version of the system consists of yes-no questions and instructions, and the patient is expected to respond non-verbally, e.g. by nodding or pointing with their fingers.

BabelDr differs from general speech translation systems like GT in several important respects. In particular, both speech recognition and translation are performed by domain-specific rule-based methods, as opposed to GT's general-purpose data-driven methods. As explained in (Rayner et al., 2016), they are for convenience split into multiple pieces, one for the source language and one for each target language, with the parts relevant to each language placed in different files; source and target languages are linked through canonical representations of the source-language utterances. The files are combined at compile-time, and the result is converted first into Synchronised Context-Free Grammar form (Aho and Ullman, 1969), and then into a GrXML grammar which can be compiled and run on the Nuance Toolkit 10.2 platform. This means that speech recognition, parsing and translation are all performed by the Nuance Toolkit engine.

At runtime, the system echoes back the canonical form of the sentence to the source-language user, only producing a translation if the source-language user approves. The canonical form thus acts both as a pivot for translation and as a back-translation to verify recognition. It was designed with the help of HUG to be the less ambiguous and the most explicit as possible, for example a sentence such as *avez-vous l'impression d'être fiévreux ?* "do you feel you're running a temperature?" is mapped to *avez-vous de la fièvre ?* "do you have a fever?". Similarly *où va la douleur ?* "which way is the pain going?" corresponds to *pouvez-vous montrer avec le doigt où irradie la douleur ?* "could you show me with your finger the direction in which the pain is radiating?".

Target-language utterances can be realised in spoken form either using the Nuance Text-to-Speech Engine (TTS), or using prerecorded multimedia files. This functionality is needed for low-resource languages like Tigrinya, which currently lack TTS engines, and also for translation into sign language (Ahmed et al., 2017). The platform is entirely web-based. The runtime system runs on a cloud server and is accessed through a thin client running on a normal web browser. Content is remotely uploaded and compiled through a web interface. The methods used were developed on previous projects and have been described elsewhere (Fuchs et al, 2012; Rayner et al, 2015).

Linguistic coverage is organised into domains, centered around body parts (abdomen, head, chest and kidneys/back); there is nontrivial overlap, since some questions are common to all domains. At the time of writing, each of the four domains has a semantic coverage of around 2000 utterance types, with an associated grammar that uses a vocabulary of about 2000–2500 words and expands to on the order of tens of millions of surface sentences. The system supports translation from French to Arabic and Spanish, and there are partial sets of translations for Tigrinya, English, LSF-CH (Swiss French sign language) and Auslan.

The Babeldr interface was designed to resemble the GT interface, but presents several important differences (both interfaces are shown in Figure 1). First, since BabelDr is a phraselator, it provides help and gives access to the list of possible canonical sentences covered by the system. After each recognition event, the list of examples is up-

Figure 1: BabelDr and GT interfaces

dated and the system automatically moves to the recognized sentence, allowing to see related questions. Second, in BabelDr input is by speech only. If the system does not recognize the utterance correctly, the user has to speak again. In GT, users can edit the recognition result by typing, or bypass speech recognition entirely and type input. Third, instead of displaying a recognition result, BabelDr displays the canonical form of the spoken utterance. Finally, the way to use the microphones differs, GT being push-and-talk and BabelDr push-and-hold.

3 Experiment

3.1 Goal

The aim of this user study is to measure the impact of the medium (BabelDr, GT) on the diagnosis made by doctors. Both systems were used by doctors at HUG or medical students to perform a medical diagnosis, based on two scenarios. For each scenario (appendicitis and cholecystitis), a patient was standardized by HUG. The two patients both received the a priori list of symptoms for the disease they present. They were instructed to give a negative or noncommittal answer for all other symptoms. The order of system and scenario versions were balanced among participants, each participant performing two diagnoses, one with BabelDr and one with GT, in an alternate order. The experiment ends when the doctor reaches a diagnosis.

3.2 Languages and domain

The language pair for the study was French into Arabic. For BabelDr, the "abdomen" domain was used. In both systems, TTS was used for speech output.

3.3 Participants

All participants were recruited at the hospital and were paid for the task:

Arabic speaking patients: two standardized Syrian patients, one male and one female.

French speaking doctors: four medical students and five doctors (clinical chiefs) working at HUG.

3.4 Location and duration

The study took place at the HUG evaluation lab and was organized in two main sessions, a pre-test with the four students and the main user study with the five doctors. One week before each test, participants received a short introduction to both systems and were given 30 minutes to practice and ask questions.

3.5 Data collected

The following data were collected during the experiments: video recordings of the room, screen capture videos, eye tracking data, diagnoses reached by doctors after each scenario, demographic and satisfaction questionnaires.

The videos and screen captures were transcribed, in particular what was said by the doctors, the recognition result and the translation into Arabic. In the following sections, we analyse these results focusing on the doctor-patient communication rather than system performance. We will therefore look mainly at interactions which reached the patients. Section 4.1 focuses on interactions sent to translation by the doctors, Section 4.2 on the quality of their translation, section 4.3 on diagnosis and section 4.4 on user satisfaction.

4 Results

4.1 Interactions with the system

Table 1 summarizes the interactions with the two systems. Overall, the number of interactions was similar for both. On average, the doctors did 30 interactions per dialog, while students have an average around 45.

Table 1: Interactions with the systems

	Total	Translated
GT		
Students	181	179 (99%)
Doctors	150	140 (93%)
BabelDr		
Students	187	128 (68%)
Doctors	156	109 (70%)

Since the two systems function differently both in terms of recognition and translation, as described in Section 2, the definition of a successful interaction with each of the systems is not straightforward. Since the source language users do not understand the target language, they can only judge the correctness of speech recognition. In this section, we consider accepted interactions, namely those where the user has found the recognition result to be satisfactory, and has validated this either by sending the utterance to translation (in BabelDr) or by oralizing the translation (in GT, where translation is enabled by default). This does not necessarily imply that the recognition result exactly matches the spoken utterance, but rather that it expresses the meaning intended by the user.

Table 1 shows that the number of interactions sent to translation and oralized for the patient is higher in GT than in BabelDr, with 99% (students) and 93% (doctors) of interactions accepted *vs* 68% and 70% in BabelDr.

Table 2: BabelDr: non oralized interactions

	Students	Doctors
1. Out of coverage	**39**	**33**
a. Out of domain	16	8
b. Out of grammar	23	25
2. In coverage	**15**	**11**
a. Canonical rejected	3	6
b. Recognition error	12	5
3. Interaction issues	**5**	**3**
Total not translated	**59**	**47**

A closer analysis of the rejected interactions in BabelDr shows different causes. These interactions are detailed in Table 2. About two thirds of rejected interactions are cases where the user produced an utterance that was not covered by the system (1). These can be split into two types. First, interactions that were not among the canonical sentences included in the domain coverage of the system (1a). These were mostly wh-questions (*quel est votre problème ? "what is your problem?"*) and declarative sentences that were not part of the usual anamnesis questions (*je vais appeler l'infirmière "I will call the nurse"*). Second, interactions using surface forms not covered by the grammar (1b). This accounts for 23 of the student's and 25 of the doctor's interactions. They are due either to gaps in the coverage or to users not complying with the instructions. Although participants were instructed not to use ellipsis, coordination, complex sentences or informal language during the introduction, some used them anyway, often resulting in incorrect recognition results. This category also includes disfluencies.

A second group (2) includes failed interactions for in coverage utterances. Some were rejected because users did not find the canonical appropriate (2a), or decided to ask something else instead. The rest (2b) were caused by recognition errors, sometimes due to a long silence at the beginning of the interaction.

Finally, for a small number of cases, interaction issues led to bad recording or translation (3).

Another aspect to consider when comparing the number of rejected interactions is that BabelDr only allows speech input while in GT, participants could also type when recognition did not work. We observe that the students corrected or typed 50 (28%) and the doctors 5 (3%) of their GT interactions. Table 3 shows the detail of these interactions. Between 2 and 3% of GT recognition results were corrected manually (for example, *allez-vous à* **sel** *normalement* "can you go to the **salt** normally" –> *allez-vous à* **selles** *normalement* "can you go to the **bathroom** normally; *est-ce que la couleur* **et** *brune* "**and** the colour brown" –> *est-ce que la douleur* **est** *brune* "**is** the colour brown"). For the students, we also observe a larger number of cases where modifications were related to incorrect interactions with the system, e.g. where they forgot to stop the microphone after the interaction. Finally, also for the students, we have a number of cases where users preferred typing input rather

than speaking (12%).

Table 3: Interactions modified by typing in GT

	Students	Doctors
Correction of rec. result	6 (3%)	3 (2%)
Bad interaction	22 (12%)	2 (1%)
Typed input only	22 (12%)	-
Total	50	5

4.2 Translation quality

The sentences sent to translation by doctors (canonical form for BabelDr, recognition result or typed input for GT) were evaluated in terms of adequacy and comprehensibility by three Arabic advanced translation students of the Faculty of Translation and Interpreting of Geneva University. Adequacy was annotated on a four point scale (nonsense/mistranslation/ambiguous/correct) and comprehensibility on a four point scale (incomprehensible/syntax errors/non idiomatic/fluent). Evaluation was carried out in context and taking into account the sex of the patient (male or female). Table 4 presents the majority judgements for adequacy and comprehensibility as well as the number of cases where no majority was reached. Inter-annotator agreement for both evaluations is moderate (Light's Kappa for adequacy: 0.483; for comprehensibility: 0.44) according to (Landis and Koch, 1997).

Table 4: Translation evaluation (doctor's interactions sent to translation)

	BabelDr		**GT**	
Adequacy				
no majority	1	1%	10	7%
nonsense	0	0%	53	38%
mistranslation	0	0%	0	0%
ambiguous	7	6%	24	17%
correct	101	93%	53	38%
Total	109		140	
Comprehensibility				
no majority	0	0%	14	10%
incomprehensible	0	0%	52	37%
syntax errors	3	3%	18	13%
non idiomatic	3	3%	3	2%
fluent	103	94%	53	38%
Total	109		140	

We observe that GT is less adequate and comprehensible than BabelDr. The evaluators also fail to reach a majority judgement more often for GT than BabelDr, suggesting that these translations are more difficult to evaluate. Interestingly, the BabelDr translations are not always considered as correct. In BabelDr, translations account for the gender of the patient, but were intended to be neutral in respect to cultural, educational and formality aspects. Evaluators disagree with some translation choices. An interesting improvement of the system would include more different patient profiles. The translators were also strict, marking some BabelDr translations as incorrect although they were completely meaningful (for example, *pouvez-vous me montrer avec le doigt où est la douleur ?* "could you indicate with your finger the pain location ?" –> هل يمكنك وضع اليد على منطقة الألم ؟ "could you indicate with your hand the pain location ?". This shows the subjectivity of human evaluation and the need to give better evaluation guidelines and to focus more on oral comprehension in future evaluations.

4.3 Diagnosis

Each of the 9 subjects had to find 2 diagnoses (1 appendicitis and 1 cholecysticis), one with each system. With GT, 5/9 doctors found the correct diagnosis, against 8/9 with BabelDr, which suggests that BabelDr is more suitable for the diagnostic task. If we look at the doctors only, they all found the right diagnosis with BabelDr, against 4/5 with GT. These results suggest that, even if it is possible to reach a correct diagnosis with bad translations, correct translations facilitate the task. It is interesting to see that BabelDr seems to help students to perform better diagnoses (1/4 diagnoses correct with GT and 3/4 with BabelDr), perhaps because the system gives access to the list of canonical sentences and helps them ask relevant questions.

4.4 Satisfaction

At the end of the task, doctors completed a questionnaire which confirms the quantitative results. Even if they felt constrained with both systems, they agreed that with Babeldr:

- they could ask enough questions to be sure about the diagnosis (only 1/9 negative opinion with babelDr vs 4/9 in GT)
- they were confident in the translation to the target language (1/9 negative opinion with BabelDr vs 8/9 in GT)

- they liked the way recognition results are presented (0/9 negative opinion vs 3/9 with GT).

- they could integrate the system in their everyday medical practice (1/9 negative opinion with BabelDr vs 5/9 with GT)

The participants had the same subjective perception of recognition quality with both systems. 3/9 participants think that they couldn't be recognized easily and 4/9 think they could. Others were neutral. In the post-experiment interviews, the doctors often mentioned the difficulty of expressing their questions as yes/no questions, as this is unusual in the anamnesis dialogue.

5 Conclusion

The data collected in this user study show that despite a very good speech recognition component, GT's translations are far less adequate and less comprehensible than BabelDr's. Along with the lower confidence expressed by the doctors in this system, this suggests that GT is not precise enough for the task, corroborating our first hypothesis. Despite this, GT allows some users, mainly the doctors, to reach a correct diagnosis. However, correct diagnoses were far more frequent with BabelDr.

This study has also provided insights into the suitability of a limited coverage phraselator such as BabelDr for this task. Although we observe more rejected interactions than for GT due to the rule-based approach, this was not perceived as particularly limiting by the users, who felt they could ask enough questions. This suggests that BabelDr is a promising tool for the task. Future enhancements of the system include training of a statistical recognizer and implementation of robust matching methods to reduce the number of failed speech interactions.

This experiment allowed us to collect a corpus of 18 diagnostic dialogues performed with two different tools, which can be used to study many different aspects of doctor-patient communication or for a shared task.

6 Acknowledgements

This project is financed by the "Fondation Privée des Hôpitaux Universitaires de Genève". We thank Manny Rayner for his comments on this paper.

References

Ahmed, F., Bouillon P., Destefano, C., Gerlach, J., Hooper, A., Rayner, M., Strasly, I., Tsourakis, N. and Weiss, C. 2017. Rapid Construction of a Web-Enabled Medical Speech to Sign Language Translator Using Recorded Video. to appear in Future and Emergent Trends in Language Technology. Seville (Spain) - 2016, Springer.

Aho, A.V. and Ullman, J.D. 1969. Properties of syntax directed translations. *Journal of Computer and System Sciences 3*, 3:3:319–334.

Flores, Gl. et al. 2003. Errors in medical interpretation and their potential clinical consequences in pediatric encounters. *Pediatrics*.

Fuchs, M., Tsourakis, N. and Rayner, M. 2006. A Scalable Architecture For Web Deployment of Spoken Dialogue Systems. *Proceedings of LREC 2012*, Istanbul, Turkey.

Hacker, K., Anies, M., Folb, B.L. and Zallman, L. 2015. Barriers to health care for undocumented immigrants: a literature review *Risk Management Healthcare Policy*, 156:1108–1113.

Jacobs, E.A., Sadowski, L.S. and Rathouz, P.J. 2003. The impact of an enhanced interpreter service intervention on hospital costs and patient satisfaction. *Journal of General Internal Medicine*, 22(2):306–311.

Landis, J.R. and Koch, G.G. 1997. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159-174.

Patil S. and Davies P. 2014. Use of Google Translate in medical communication: evaluation of accuracy *BMJ*, 349.

Priebe, S., Sandhu, S., Dias, S et al. 2011. Good practice in health care for migrants: views and experiences of care professionals in 16 European countries. *BMC Public Health*, 11.

Rayner, M., Baur, C., Chua, C., Bouillon, P. and Tsourakis, N. 2015. Helping Non-Expert Users Develop Online Spoken CALL Courses. *Proceedings of the Sixth SLaTE Workshop*, Leipzig, Germany.

Rayner, M., Armando, A., Bouillon, P., Ebling, S., Gerlach, J., Halimi, S. and Tsourakis, N. 2016. Helping Domain Experts Build Phrasal Speech Translation Systems. José F. Quesada et al.. Future and Emergent Trends in Language Technology. Seville (Spain) - 2015, Springer, 2016:41-52.

Wassermann, W. et al. 2014. Identifying and Preventing Medical Errors in Patients with Limited English Proficiency: Key Findings and Tools for the Field. *Journal for Healthcare Quality*.

Wu A.C., Leventhal J.M., Ortiz J., Gonzales E.E. and Forsyth B. 2006. The interpreter as cultural educator of residents. *Archive of Pediatric and Adolescent Medicine 160, 1145–1150*, 160:1145-1150.