BabelDr vs. Google Translate: translation in a pharmacy setting

Mali, Rebeka Karola

UNIVERSITÉ DE GENÈVE
FACULTY OF TRANSLATION AND INTERPRETING

REBEKA MALI

BabelDr vs. Google Translate: translation in a pharmacy setting

Directrice : Pierrette Bouillon
Juré-e-s : Marie Paule Schneider Voirol, Hervé Spechbach

Mémoire présenté à la Faculté de traduction et d'interprétation pour l'obtention de la Maîtrise universitaire en traitement informatique multilingue.

Université de Genève
Janvier 2023

Acknowledgements

I could not have undertaken this journey without the guidance of my thesis director, Professor Pierrette Bouillon. She has always welcomed me with open arms, a kind smile, and gentle words in her office. She has given me all the inspiration needed to write this thesis: her hard work and outstanding research have taught me everything I know. In addition, she has reviewed my work during weekends and holidays. I cannot imagine having a better thesis director. *Merci beaucoup pour tout, vraiment.*

Words cannot express my gratitude to Professor Marie Paule Schneider Voirol. This project aimed to help pharmacies in their practice. I am beyond proud to have had the opportunity to do this project with her; without her kindness, interest, and contribution, I could never have conducted this research. I hope we have the chance to work together very soon. *Merci beaucoup.* I'm incredibly grateful to have worked with Doctor Hervé Spechbach. Since our first meeting, he has actively participated in the project, shown enthusiasm, and reviewed this work from the beginning. I am honored to have you on my jury. I hope this will be just the beginning of our many projects together. *Merci beaucoup.*

I want to express my deepest gratitude to all the participants in my study. The pharmacists for all the knowledge they have shared with me and who have been at the core of this research. I hope to have contributed to helping you with your daily work. A big thank you goes to all the participants who have played the role of patients for their incredible availability, kindness, and patience in participating in this study. I want to extend my sincere thanks to the translation students and the two pharmacists who took part in the evaluation sessions for their professionalism and insights. Without them, I could not have achieved any of these results. *Merci à toutes et à tous.*

My warmest thanks go to all my friends and work colleagues who have been around me this last year and who I consider my family. Although they have not actively participated in the study, only their presence has helped me complete this work. Thank you for listening, being there for me, and all the patience you demonstrated in my darkest days. I have grown a lot thanks to you. *Grazie, merci, and gracias.*

Last but not definitely not least, I would like to express my deepest gratitude to my fiancé, Jonathan. I'm incredibly grateful to have you in my life. Without your immense patience, corrections, advice, and the nights spent debating, this thesis would not come near to what it is. *Grazie per rendermi ogni giorno una persona migliore. Sempre.*

Contents

Chapter 1

Introduction

Language barriers decrease the quality of care [Bartlett et al., 2008]. Proper communication between patient and health care provider gives numerous advantages: reaching the accurate diagnosis, choosing the appropriate treatment, being sure that the treatment is correctly administered and adhered to, and taking the correct action when an adverse event presents itself [Vincent and Coulter, 2002]. Furthermore, patients receiving complete information about the harms and benefits of treatment are more likely to adhere to it, leading to better outcomes [Vincent and Coulter, 2002]. One study argued that patients with language barriers, and especially migrants, are three times more likely to be at risk of preventable adverse events ("an adverse event is an unintended injury or complication caused by delivery of clinical care rather than by the patient's condition")[Bartlett et al., 2008]. The Institute for Healthcare Advancement in the United States calculated an annual waste of around 73 billion dollars in 2003 due to communication issues in healthcare. The leading cause of poor communication with migrant patients is a lack of qualified interpretation [Bagchi et al., 2011]. This results in medical mistakes, less treatment adherence, and more frequent use of emergency services [Bagchi et al., 2011].

Professional face-to-face interpreting can be provided. However, scheduling conflicts, privacy concerns, and significant fees are involved. Decision-makers consider professional interpreting services expensive. For instance, the 2013–14 Refugee and Humanitarian Program received 54.3 million dollars from the Australian government's federal budget for translation and interpreting services [Chang et al., 2014]. Other affordable options are frequently used to meet this pressing need for interpreters. Telephone interpreting services are one of these, and while they are easier to organize, they are still quite expensive. Another practical solution is to turn to bilingual professionals. However, if one is working with a small minority language, this may not be a viable solution [Halimi and Bouillon, 2019]. Health practitioners also rely on patient's relatives and friends, even though studies demonstrate that doing so leads to poor communication and confidentiality breaches [Diamond et al., 2009].

Machine translation (MT) is another viable alternative to human interpretation. Recent research reveals that Google Translate (GT) is being used in healthcare settings for obvious reasons such as convenience and low-cost [Taylor et al., 2015]. Despite the increasing employment of GT in healthcare contexts, research suggests

that doctors are skeptical of using broad-coverage speech translation systems to overcome language obstacles in a medical environment, including diagnosis, for which these systems have not been trained [Bouillon and Spechbach, 2016]. The usage of GT cannot be addressed without raising significant concerns about medical risks to patient life and ethical concerns about processing patient information in a widely utilized system that stores that information in its server [Wade, 2011]. Moreover, studies have shown that this tool is not reliable in medical settings: accuracy for African languages is very low (45%) as well as for Asian languages (46%) [Patil and Davies, 2014]. Furthermore, Google Translate also poses ethical problems regarding the collection of personal data and is incompatible with the Swiss Data Protection Law [Spechbach et al., 2019]. The research presented here aims to find a solution that yields reliable output and combines speech recognition with generative language: BabelDr. This flexible phraselator employs speech recognition and is currently used at the emergency department at the Geneva University Hospitals to communicate with allophone patients [Bouillon and Spechbach, 2016]. The tool is intended to help out refugees, foreign-languages speaking residents, deaf and hard of hearing.

1.1 Context and motivation

In light of the forthcoming shortage of family physicians in Switzerland, pharmacists, as medically trained specialists, should occupy a more important role in primary medical care. The Federal Council of 30/11/2012 believes that "cooperation between the various health professionals would not only help to remedy regional shortages in health care providers but would also be an essential means of ensuring safe and integrated care for the benefit of the patient"[Ruth, 2012]. Therefore, the Federal Council has taken all the necessary steps to make the best use of pharmacists' skills in primary health care [Ruth, 2012]. One possibility to lower the burden of emergency rooms is performing triage for minor health disorders in pharmacies [Stampfli et al., 2021]. Thanks to triage in the pharmacy, the patient's risk is assessed, the medical needs can be prioritized, and the proper care stream is indicated to the patient. A significant advantage of triage performed by pharmacists is that care in pharmacies is available without prior appointment and for long working hours [Stampfli et al., 2021]. One study in the UK has shown that pharmacy triaging has reduced the workload of physicians when faced with minor health problems, with 68–94% of advised clients reporting that they resolved their medical needs [Stampfli et al., 2021].

In Switzerland, according to the 2019 revision of the Federal Act on Medicinal Products and Medical Devices, pharmacists can dispense, within a well-defined framework, prescription medication without the advice of a physician. Not only in cases of urgency but also for several drug indications and treatments established by the Federal Council [Confédération, accessed January 8, 2023]. Pharmasuisse, the Swiss association of pharmacists, to promote this new service provides pharmacists with decision trees on minor health disorders, which can be used to triage, guide the patient with self-medication or self-care, or address the latter to the proper care system. This contributes immensely to lowering the burden on the emergency

departments and the whole healthcare system [Stampfli et al., 2021]. These decision trees are accessible via the service "Netcare," both via web application and print-based forms, enabling pharmacists to perform triage using scientific algorithms developed in collaboration with physicians and pharmacists. They include a "structured assessment of the symptoms, urgent red flags warranting immediate referral, suggested triage outcomes, and treatment recommendation" [Stampfli et al., 2021]. A Swiss analysis discovered that 641 (15.1%) clients did not have a general practitioner. Nowadays, pharmacies can take care of patients without established access to health care [Stampfli et al., 2021].

However, in the context of the current European refugee crisis, not only are there more and more patients seeking medical help, but pharmacists are also often faced with foreign-language speaking residents or refugees who seek their help but cannot communicate the problem. Only for the last year (from 1 January 2022 to 30 November 2022) has Switzerland provisionally admitted 61'553 asylum seekers. Most of them come from Europe (60'588 of which 59'738 from Ukraine), followed by Asia (730) and Africa (187) [Confédération, accessed December 30, 2022]. So far, to our knowledge, no technology has been developed to help pharmacists overcome the language barrier. It would be ethically correct that all patients were offered the same quality of care, with the possibility of interacting with health care professionals [Spechbach et al., 2019]. Asking ad hoc interpreters such as parents, family members, and social workers may result in errors of clinical consequence such as omissions about drug allergies, errors in the instructions on the duration, the dose, and the frequency of drug medication [Flores et al., 2003].

1.2 Objectives and research questions

The research aims to assess whether BabelDr, a speech-enabled fixed-phrase translator currently used at the Geneva University Hospitals (HUG), could assist pharmacists in triaging and making the correct diagnosis when faced with Arabic-speaking patients.

More in detail, this study aims to answer the following question:

Is BabelDr better suited than Google Translate in a pharmacy setting to triage Arabic-speaking patients?

The research has a quadruple objective:

1. Assess which system between BabelDr and Google Translate is the best for performing triage with Arabic-speaking patients in a pharmacy setting;

2. Evaluate which system between BabelDr and Google Translate is the most usable in terms of successful interactions and time for performing triage in a pharmacy setting;

3. Rate whether the translations with BabelDr and Google Translate are accurate, fluent, or dangerous from the point of view of patients who know little

standard Arabic and speak different vernacular Arabic dialects such as Moroccan, Tunisian, and Egyptian;

4. Rate whether the translations with BabelDr and Google Translate are accurate or dangerous from the point of view of a pharmacist (domain expert).

To each objective corresponds one or more research hypotheses which are resumed in table 1.1

Arabic has been chosen as the language for this research as it represents one of the pressing needs in the medical setting in Geneva, especially for refugees [HUG, 2022]. We have investigated whether BabelDr and Google Translate are suitable for standard Arabic during a pharmacy-based triage and how well both systems perform in cases where pharmacists face patients who have little understanding of standard Arabic and speak vernacular Arabic dialects such as Tunisian, Moroccan, and Egyptian. The reasons for this will be better explained in section 3.3.

Objective	Hypothesis
1. Asses which system between BabelDr and Google Translate is the best for performing triage with Arabic-speaking patients in a pharmacy setting	Pharmacists are always able to reach a correct diagnosis with BabelDr and Google Translate.
1. Assess which system between BabelDr and Google Translate is the best for performing triage with Arabic-speaking patients in a pharmacy setting	Pharmacists prefer employing BabelDr rather than Google Translate to perform triage as BabelDr has been specifically developed for the medical sector.
2. Evaluate which system between BabelDr and Google Translate is the most usable in terms of successful interactions and time for performing triage in a pharmacy setting	BabelDr is the most usable system to perform triage in pharmacies compared to Google Translate with regards to successful interactions due to the speech recognition errors of the latter; with regards to time, as Google Translate is widely known and used, we hypothesize that pharmacists will reach a correct diagnosis faster with this system rather than with BabelDr.
3. Rate whether the translations with BabelDr and Google Translate are accurate, fluent, or dangerous from the point of view of patients who know little standard Arabic and speak different vernacular Arabic dialects such as Moroccan, Tunisian, and Egyptian	Since both systems employ standard Arabic, only a small percentage of the questions will be accurate, comprehensibility scores will be low, and at least 10% of the translations will be dangerous both with BabelDr and Google Translate.
4. Rate whether the translations with BabelDr and Google Translate are accurate or dangerous from the point of view of a pharmacist (domain expert)	From a pharmacist (domain expert) point of view, all of BabelDr's translations are accurate, and none are dangerous. Nevertheless, not all of Google Translate's translations are accurate, and at least 3% are dangerous.

Table 1.1: Research objectives with their respective hypothesis.

1.3 Methodology

In order to provide an answer to these hypotheses, we set up an experiment that allowed us to collect quantitative and qualitative data in different steps, namely:

1. We invited seven pharmacists from Pharma24 and one from another pharmacy to perform tests with both systems: BabelDr and Google Translate; we also recruited Arabic-speaking participants to play the role of the patients.

2. Pharmacists were asked to perform two triages based on different scenarios, one with each system.

3. The system usage was analyzed in terms of successful and failed interactions, and the time required to reach a diagnosis was calculated in order to determine which system is more suitable in a pharmacy setting.

4. Pharmacists' satisfaction questionnaires were collected to understand which system they preferred.

5. Their triage questions were gathered with the respective translations. We asked three master-level translation students and two pharmacists to evaluate the Arabic translations based on accuracy, fluency, and dangerousness. In particular, students were asked to evaluate the translation from the point of view of someone who speaks a particular Arabic vernacular dialect (Tunisian, Moroccan, and Egyptian) and with a little understanding of Standard Arabic to give a more profound layer of analysis to the study by taking into account accessibility; the pharmacists were asked to evaluate the translations from an expert's point of view with standard Arabic.

6. Both translators' and experts' inter-agreement has been calculated: Light's Kappa for the translators and Cohen's Kappa for the experts have been used as units of measure to calculate the reliability of the evaluation's results.

1.4 Structure

This thesis can be divided into three main sections: the first aims to introduce the state of the art in the domain, the second to present the tests and evaluations conducted, and the third to illustrate the results.

Chapter 2 will focus on the state of the art in the domain; it will have different sections introducing communication problems in health settings and the different solutions employed so far to overcome the language barrier.

Chapter 3 will be centered on the methodology applied to answer the research question and its relative hypotheses. Namely, it will present in detail the objectives and research hypotheses, the system settings of the applications employed for the experience. It will then focus on the the tests and evaluations procedures. It will lastly illustrate the quantitative and qualitative data collected.

Chapter 4 will present the study's results, following the methodology's different points. It will reveal the demographic questionnaires' results, the diagnoses, the system usage (interactions with the system and time), and the users' satisfaction; lastly, it will offer the results of the translators' and experts' evaluation of the translations with both systems.

Chapter 5 will draw a link between the research questions and the answers obtained.

Chapter 6 will conclude by presenting a summary of the study and by proposing new perspectives for future work and research to overcome some of the limitations encountered.

Chapter 2

State of the art

This chapter will be divided into two main sections. In the first part, we will introduce language barriers first in the pharmacy context (2.1) and subsequently in the medical context (2.2). In the second part, we will present the solutions (2.3) to the language barriers, such as phraselators (2.3.1) and machine translation (2.3.2). We will give a few examples of phraselators such as BabelDr (2.3.1.1), CALD Assist App (2.3), MediBabble (2.3.1.3), and UniversalPharmacist Speaker (2.3.1.4). We will also suggest another solution by introducing Google Translate (2.3.2.1). Lastly, we will summarise two studies that compare BabelDr and Google Translate: Gerlach (2022) and Bouillon (2017) (2.3.3).

2.1 Language barriers in pharmacies

As we have seen in our introduction (1), a significant concern in migrant populations is medication safety due to adverse drug events and an increased risk for interactions [Schwappach et al., 2012]. Various cases are reported in the literature where misunderstandings due to language barriers led to children overdosing. For example, it was the case of a 10-month-old girl with an iron-deficiency anemia who was hospitalized due to an overdose of iron after her non-English speaking parents were given an English prescription and medication instructions which were misinterpreted [Bradshaw et al., 2007]. In another case, a 6-week-old boy was hospitalized due to a barbiturate overdose caused by a medication dosing error by the not English-speaking mother who did not understand the dosing instructions given only in English [Bradshaw et al., 2007].

In 2011, the World Health Organization published the guidelines on good pharmacy practices, stating that "pharmacists should acknowledge unique patient considerations such as education level, cultural beliefs, literacy, native language and physical and mental capacity in all individual patient assessments" [Organization, 2011]. However, we need to determine to what extent pharmacies achieve this goal [Schwappach et al., 2012]. There is not much evidence regarding the safety of delivering pharmaceutical care for migrants, despite a large migrant population in European countries [Schwappach et al., 2012]. For example, a survey of medical services in Switzerland regarding language barriers in health care suggested that communication with allophone patients is perceived as significantly difficult by the Swiss medical service [Bischoff et al., 1999].

A quantitative survey commissioned by the University of Neuchâtel by the Swiss Federal Office for Migration studied the information channels of Albanian and Turkish speakers [Dahinden et al., 2009]. What emerged was that pharmacies were frequently mentioned by women and young people as places of information since they do not need an appointment and that the services and advice of pharmacists are more affordable than a medical visit [Dahinden et al., 2009]. Furthermore, the cited advantages were the pharmacies' anonymity, informality, and accessibility [Dahinden et al., 2009]. As in most European countries, there has never been an empirical assessment of the safety and quality of pharmaceuticals provided by Swiss public pharmacies to migrants [Schwappach et al., 2012]. One study, whose sample included all heads and owners of public pharmacies in Switzerland who are members of the Swiss Pharmacist Association (about 75% of all public pharmacies), reported that about 10% of pharmacies fail at least once a week to explain the drug therapy to foreign-language patients. 64.7 % declared that the problems are mainly because of language barriers they encounter when counseling medication to non-Swiss patients. 25.3 % reported a combination of cultural and language barriers [Schwappach et al., 2012]. Many pharmacists declared that when foreign-language patients leave their pharmacy, they are concerned about medication safety; 14.0 % reported that this occurs at least monthly and 8.5 % at least weekly [Schwappach et al., 2012].

One way to overcome language and cultural barriers is to employ multilingual staff; pharmacies adopted this primary strategy to improve the quality of pharmaceutical counseling [Schwappach et al., 2012]. However, pharmacies mostly rely on pharmacy assistants, often second or third-generation migrants, to offer medication counseling [Schwappach et al., 2012]. It is still to be determined whether pharmacy assistants are adequately trained to provide medication counseling in various languages. This should be ensured and evaluated by pharmacists [Schwappach et al., 2012].

In the US, a study [Bradshaw et al., 2007] conducted among public pharmacies in Milwaukee County reported that out of 128 pharmacies, 47% only sometimes/never can print non–English-language prescription labels, and 64% only sometimes/never can verbally communicate in non-English languages. Of those capable of doing so, 95% frequently use bilingual staff or a computer program for prescription labels. In contrast, for verbal communication, around one-third claim to use telephone interpretation services, and two-thirds claim to use bilingual staff and other methods [Bradshaw et al., 2007].

A recent study in Australia [Mohammad et al., 2021] conducted focus group discussions with 30 pharmacists recruited from metropolitan Sydney. Researchers found that in a community pharmacy setting, the major barrier to competent care is the language proficiency between patient and pharmacist. Pharmacists were aware and mindful that culturally and linguistically diverse (CALD) patients with particular religious affiliations or cultural backgrounds might prefer consulting a same-gendered pharmacist or assistant for gender-specific health issues [Mohammad et al., 2021]. Participants in the study were also aware of particular dietary restrictions of CALD patients, such as the choice to consume vegan or halal medicinal products;

these were reported as being the predominant belief-related demand [Mohammad et al., 2021]. Pharmacists expressed issues with their professional satisfaction: some said that they felt like they were not doing their job as they ought to as a health professional; others reported feeling unease by the fact that children are reading prescription labels to their parents and that in some cases it is the pharmacists who are liable for these matters; some reported being ashamed of not asking confirmation if the CALD patients understand, while others felt concerned about the fact that they had no way of knowing whether the medication they were dispensing was safe or appropriate for them since they could not effectively communicate [Mohammad et al., 2021]. Participants also reported means by which they could ensure culturally competent care. One of these was increasing awareness of existing resources through educational courses or a translation service. However, participants expressed having to wait to talk to someone when needing an interpreter fast [Mohammad et al., 2021]. Other means often mentioned would be to offer multilingual labels given their straightforward character: there patients can find all they need and want (i.e., what the medication is for and how to assume it) [Mohammad et al., 2021]. In the same study, some pharmacists mentioned having used Google Translate. However, they expressed some concern about the tool, given that a health professional body did not officially recognize it. Professional translators were not considered relevant resources in a busy community pharmacy due to the need to wait for them. Participants also expressed the need to receive professional training on how better serve CALD patients [Mohammad et al., 2021].

2.2 Language barriers in the medical context

Cohen and Flores [Cohen et al., 2005] published an article arguing that during pediatric hospitalization in the USA, Spanish-speaking patients whose families have a language barrier seem to have an increased risk for serious medical events compared to families who do not experience language barriers. The authors hypothesize that some healthcare professionals believe in having a solid command of the Spanish language when in reality, they have not. This entails that fewer interpreters are called for these patients, and poor communication can lead to adverse outcomes and medical errors [Cohen et al., 2005].

To overcome the language barrier, youngsters (nine to 18 years old) are sometimes used as interpreters by choice: either because the family trusts the kid's language skills or because the kid knows their parent's illness and how it affected their life [Free et al., 2003]. On the one hand, young people enjoy being able to help family members and also be able to take on a responsible role by demonstrating their language skills. On the other hand, this hard work might be time-consuming and take away their play or school time. More importantly, some young people reported being frustrated and angry when they could not interpret as they wished or when they were caught in disagreements [Free et al., 2003]. Others reported feeling embarrassed by knowing their parents' sensitive information (this is the case of young men who refuse to translate women's health issues) or having to tell them what to do. Some had difficulty accepting bad news and did not want to upset their relatives by passing them on [Free et al., 2003]. Finally, a few reported being blamed when they were not understood and even being yelled at by members of the family [Free et al.,

2003]. Furthermore, there have been some cases where the parent disagreed with the physician. In this situation, the doctor assumed there was a problem with the translation rather than accepting that the patient might disagree with their advice [Free et al., 2003]. Several young people reported physicians being resentful since they could not talk directly to the patient and that the conversation was taking too long. This attitude was implicit in the doctor's angry facial expressions and disapproving looks [Free et al., 2003].

Due to cost concerns and scheduling difficulties, professional interpreters appear to be used only in the absence of other available options. The choice to use professional versus ad hoc interpreters seems to be influenced by three main factors: availability of bilingual staff, perceptions of interpreting quality, and cost concerns [Bischoff et al., 1999]. Data suggest that professional interpreters are called in only after other strategies have failed due to cost concerns and practical issues. The clinical staff is less familiar with organizing an appointment with an interpreter and less comfortable working with a non-staff interpreter [Bischoff et al., 1999]. Diamond et al. [Diamond et al., 2009] found that physicians used ad hoc interpreters even though they believed the quality of care could be compromised. They tended to normalize this practice, emphasizing that practical and time constraints limited their ability to call on professional interpreters. This indicates that family members may be the first strategy tried when bilingual staff is not available, but that bilingual staff is preferred. Their language skills may be superior to those of family members, and collaboration may be perceived as more accessible due to their medical and institutional knowledge [Bischoff and Hudelson, 2010]. There are many potential practical and financial benefits to identifying and using bilingual healthcare staff to double as interpreters. This strategy can be integrated into existing clinical routines and has fewer visible costs than professional agency interpreters. However, there are invisible costs involved with removing a staff member from one role to fulfill another, and bilingual staff should ideally receive training in interpreting, as bilingualism is insufficient to ensure adequate interpreting skills [Bischoff and Hudelson, 2010].

2.3 Solutions

In order to avoid using ad hoc interpreters, bilingual staff, or expensive translation telephone services, there might be more affordable and trustworthy solutions that can help pharmacists communicate with CALD patients.

2.3.1 Phraselators

Phraselators are machine translation systems that are based on translation memories as opposed to machine translation. They are based on translations previously completed by human translators and then entered into the system. Thus, they may prove to be a great alternative to generalist systems, especially in light of the accuracy of the suggested translations. Furthermore, phraselators offer the benefits of reliability and portability [Seligman and Dillinger, 2013]. Although these tools offer limited coverage and therefore do not solve all communication problems, recent studies demonstrate that they are often chosen over machine translation in certain

safety-critical circumstances because they are considered more trustworthy and reliable [Panayiotou et al., 2019]. In the following sections, we will introduce some of the most known phraselators employed in the medical domain.

2.3.1.1 BabelDr

BabelDr is a flexible phraselator that employs speech recognition that has been developed jointly by the Department of Multilingual Information Processing of the University of Geneva and the Geneva University Hospitals (HUG). The project aims to create a speech-to-speech translation system for emergencies that meets three requirements: reliability, portability, and data security for low-resource languages needed at the HUG. It is intended to enable French-speaking medical professionals to conduct triage and diagnostic interviews with patients who speak Tigrinya, Albanian, Arabic, Dari, Spanish, Swiss-French sign language, and Farsi [Bouillon et al., 2021]. Recently, Ukrainian, Russian, and various Arabic dialects have been added. It enables a medical expert to conduct a preliminary medical examination dialogue using a decision-tree method to identify the nature of the patient's condition and the best course of action to take [Bouillon et al., 2017].

BabelDr is a web program that works on desktop computers and mobile devices. Built on the concept of a phraselator, it uses a small number of core sentences collected from medical professionals and pre-translated by professional translators and interpreters. In addition, it has a speech recognition feature for easier usability and more natural communication with patients [Bouillon et al., 2021]. Instead of searching for phrases in menus, medical professionals can talk freely since the system will map their phrases to the closest pre-translated core sentence. In order to make sure the doctor understands precisely what is being translated for the patient, this statement is then provided for validation in a back-translation process. After that, the patient can answer using a pictogram-based interface [Bouillon et al., 2021]. Figure 2.1 from [Bouillon et al., 2021] illustrates how BabelDr functions.

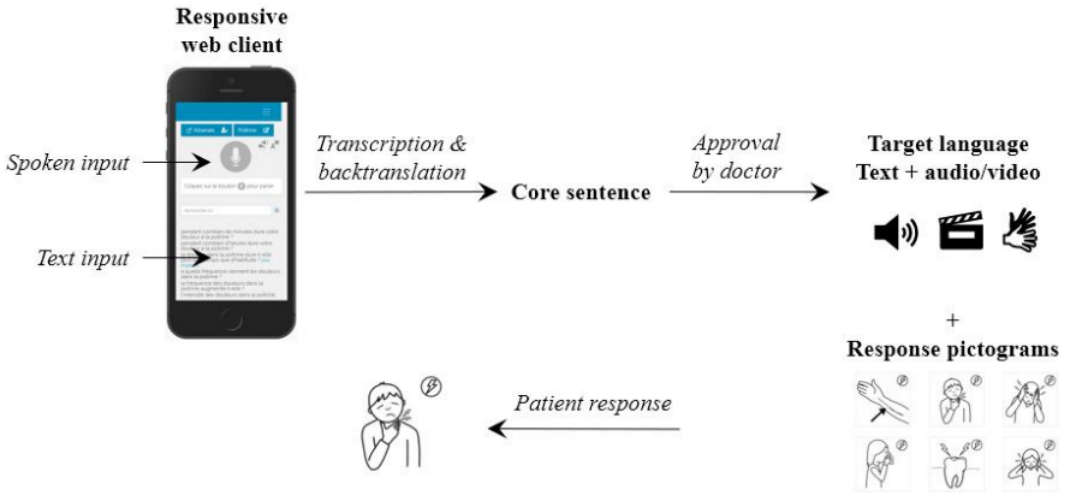

Figure 2.1: BabelDr Functionality

The back-translation step in BabelDr is crucial because it maps the voice recognition output to a core sentence that is then presented to the doctor for approval. The back-translation task can be viewed as a sentence classification task where the core sentences are the categories [Bouillon et al., 2021]. Then, employing a variety of techniques involving deep learning techniques, neural machine translation (NMT), and sentence classification, BabelDr can achieve 93.2% accuracy on core-sentence matching for transcriptions (assuming perfect speech recognition) [Mutal et al., 2020]. After the back-translation has been verified, BabelDr provides the patient with the output in the target language in both written and spoken form. These two formats are based on the human translations of the core-sentences [Bouillon et al., 2021].

A medical phraselator requires high translation quality. Thus qualified translators produce the translations. The translations are intended for individuals with no medical background and are written so patients with low literacy can understand them. Sentences are adapted to account for cultural factors, such as sensitive or intimate subjects that are not often discussed, such as sexual habits [Halimi et al., 2021]. The translators had to pick phrases that would work in both the written and spoken forms because the system offers both. According to a recent assessment of the translations for two of the system's target languages (Albanian and Arabic), these translations are simple to grasp, making the system more reliable than MT, and they are also perceived as more trustworthy [Gerlach et al., 2022]. Because audio provides several competitive advantages to patients, it has been an effective output medium for BabelDr. It lessens the strain of looking at a screen, which can be difficult in a medical context due to factors such as how the doctor and patient are positioned. It is a crucial component, especially for illiterate users, and having a system communicate to them in their local language can enhance user experience [Bouillon et al., 2021]. Although it would be feasible to have a human record every sentence that had been pre-translated, the time and expense required for recording were deemed too expensive, given the volume and repetitiveness of the words. To announce the translated questions of the physician, a Text-to-Speech (TTS) technology was chosen from the start of the project for the languages that support this technology [Bouillon et al., 2021].

The original BabelDr system was confined to yes-no questions or questions to which the patient may react nonverbally, such as by pointing to an anatomical region. Both doctors, who are accustomed to asking open-ended questions, and patients, who had limited resources to participate in the course of the conversation actively, found this constrained approach to be problematic. Various approaches to develop a bidirectional version that would enable the patient to respond in more complex ways were explored [Bouillon et al., 2021]. There are many challenges in developing a system that would allow patients to answer verbally. Many of the minority languages the system targets do not have speech recognizers, and there are few or no resources, such as speech corpora, that can be used to create such systems [Bouillon et al., 2021]. Although more straightforward to create, a text interface like those seen in conventional phraselators would not be usable by patients with low literacy. A phraselator also requires some user training to become familiar with system coverage, which is impossible for patients who come to an emergency service

[Bouillon et al., 2021]. Due to these factors, it was decided to include a primary pictograph-based response interface, as seen in Figure 2.2. Every core sentence has a collection of related response pictographs from which the patient can express their response. All patients preferred the bidirectional version, according to a task-based evaluation, because they could more effectively describe their symptoms with it [Bouillon et al., 2021].

Figure 2.2: BabelDr interfaces

The body parts (abdomen, head, chest, traumatology, habits, dermatology, and kidneys/back) organize linguistic coverage into domains. There is nontrivial overlap as some questions apply to all domains. At the time of writing, each of the six domains has a semantic coverage of roughly 2000 utterance types and an associated grammar that expands to tens of millions of surface sentences using roughly 2000–2500 words [Bouillon et al., 2017].

2.3.1.2 CALD Assist App

CALD Assist app is a mobile app used by healthcare professionals to assist with patient care when no interpreter is available. The app was developed in 2014 for clinicians and expanded in 2017 to meet the needs of the nurses. The app includes audio, images, and video content to enable communication between patient and clinician, as well as key phrases translated into ten languages [Silvera-Tawil et al.,

2021]. The app was adapted to the nursing staff because they expressed the need for it in their daily care, which includes helping the patients eat, assessing their pain, and fitting reading glasses [Silvera-Tawil et al., 2021]. Given that these interactions are frequent and short in time, interpreters might not be available, and their use is only sometimes practical [Silvera-Tawil et al., 2021]. The app has been introduced in various cities in Australia and was evaluated in terms of staff and patient acceptance, satisfaction levels, and app efficacy. The app presents a user-centric design focused on clinicians and patients. Because the nurses interact with the patients on a day-to-day basis, it presents phrases that are relevant to the nurses' needs. Two hundred commonly used sentences are translated into ten languages and arranged according to disciplines: podiatry, occupational therapy, speech pathology, dietetics, and nursing [Silvera-Tawil et al., 2021]. To facilitate communication between the clinician and the patient, each phrase is followed by answer options. Due to the user group's age, the app consists of various communication mediums to meet potential visual and audio impairments and different literacy levels. So we can find content in text, imagery, audio, and video to increase the app's utility [Silvera-Tawil et al., 2021].

Phrases are grouped based on the disciplines and follow a typical clinical interaction, starting with introduction phrases, passing through assessment or question phrases and education phrases, and closing the conversation. In figure 2.3, we can see how the app works: once a sentence has been selected, its translation appears in a large font, with the English phrase appearing above the translation in a smaller font. Underneath, we can find images or videos connected to the phrase. In the menu options, we can see that the clinician can play prerecorded audio of the phrases. In addition, the patients are offered answer options and the ability to ask follow-up questions. The key advantage of this app is that it can collect detailed information from patients via a multimodal, two-way communication [Silvera-Tawil et al., 2021].

The app can be downloaded in the Google Play and Apple App stores after six months of trial in a healthcare network in Australia [Silvera-Tawil et al., 2021]. When nurses were asked if the app was helpful, 28 out of 30 (93%) reported that it was helpful with non-English speaking patients [Silvera-Tawil et al., 2021]. The most frequent comments about the app were that the nurses would have liked more languages, phrases, images, and a feature that allowed them to type any phrase in the app to be translated. They also would have liked louder audio and the iPad to be at the patient's bedside [Silvera-Tawil et al., 2021]. Moreover, they mentioned that they needed less assistance from bilingual colleagues and family members when using the app. Therefore, they would instead use the app as the first resource, but if they needed additional help, they would ask for the help of interpreters or family members [Silvera-Tawil et al., 2021]. Regarding the patients, 6 out of 7 (86%) reported that the app was helpful by assisting them both in understanding the nurse and being understood by the nurses [Silvera-Tawil et al., 2021]. Patients' comments included the need for new phrases such as "I am cold," "I need your help," and "I am hungry/thirsty." Mainly, they asked for phrases that allowed them to explain where the pain is located and how to describe it [Silvera-Tawil et al., 2021]. Patients reported feeling more included when they used the app [Silvera-Tawil et al., 2021]. However, patients with cognitive impairment found it complex to understand long

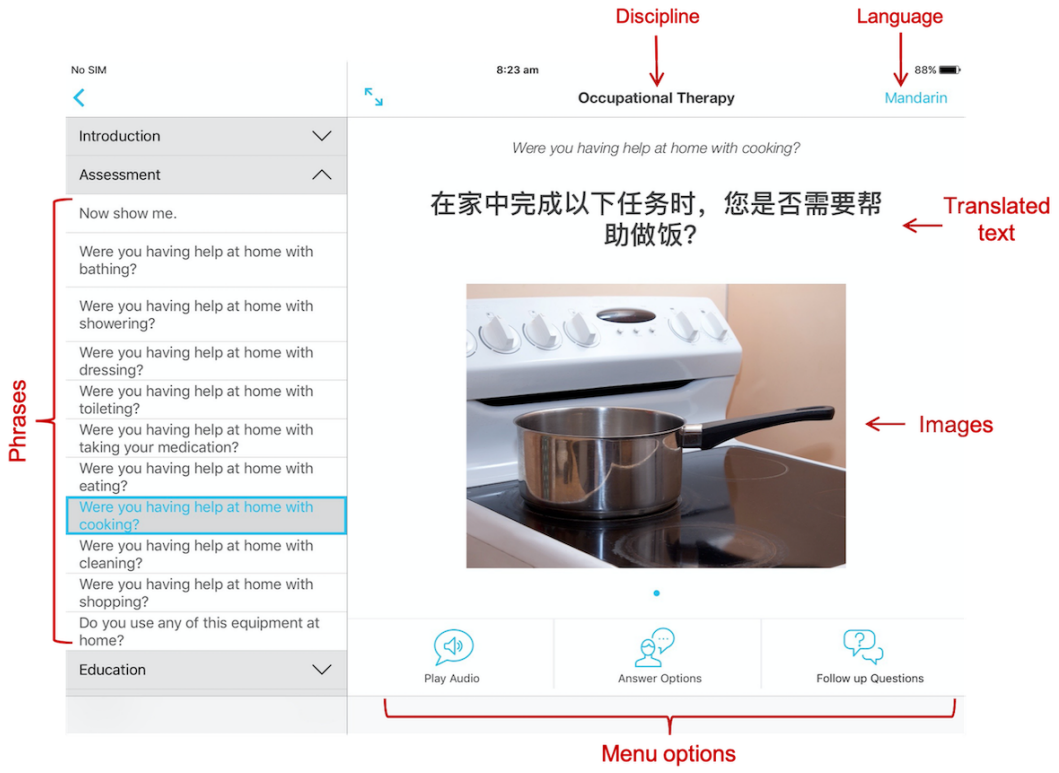

Figure 2.3: CALD app interface screenshot

sentences [Silvera-Tawil et al., 2021]. Participants said the app facilitated some exchanges, which would have been difficult without it. In one case, it helped change a patient's pain medication dosage; in another case, it helped nurses understand that a patient had pain located in the chest [Silvera-Tawil et al., 2021].

The log data analysis of app use revealed that the most popular category was pain-related sentences, and the most popular feature was playing the audio [Silvera-Tawil et al., 2021]. Participants found it helpful to have the translation written, especially with patients suffering from hearing impairment [Silvera-Tawil et al., 2021]. However, a lacking functionality of the app was being unable to slow down the audio when the spoken phrases were too fast. Interestingly, a need for the ability to translate spoken text was mentioned: they would have liked a functionality like Google Translate, which enables nurses and patients to translate speech from and into the patient's language [Silvera-Tawil et al., 2021]. Regarding the voices the app provides, the current version offers one voice per language. However, during the app's design, having both male and female voices was suggested to adapt to the patient's life experiences (i.e., a male voice might be uncomfortable for a female patient with a history of sexual assault) [Silvera-Tawil et al., 2021].

2.3.1.3 MediBabble

MediBabble is a translator app developed at the University of California by two medical students, Brad Cohn and Alex Bau [Daly, 2014]. The two students have described it as a "history-taking and physical exam application designed to improve the safety, efficiency, and quality of care for non-English-speaking patients" [Daly,

2014]. The app is organized by symptoms and contains thousands of translated questions and symptoms. Questions can be answered with gestures, a simple yes or no [Daly, 2014]. The app is free to download only on the App store of iPad, iPod, and iPhone. It does not need an Internet connection. Unfortunately, the application does not have voice recognition, so healthcare professionals either look through the list of sentences available or tap their questions via the search function. Figure 2.4 shows the main MediBabble interface:

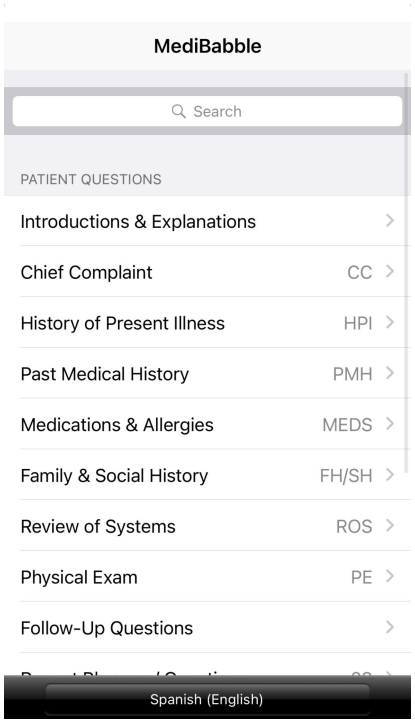

Figure 2.4: MediBabble app main interface

We can see the different main sections in which the questions are divided: Chief Complaint, History of Present Illness, Past Medical History, Medications Allergies, Family and Social History, Review of Systems, Physical Exam, and Follow-Up Questions. Once the user has selected the chosen section, a new page opens with the set of questions available, as shown in figure 2.5. When the doctor selects a question, an audio voice utters it and presents it in written form.

The program translates into six languages: English, Mandarin, Spanish, Haitian Creole, Russian, and Cantonese [Sheik-Ali et al., 2016].

In a case study that aimed to compare BabelDr and MediBabble, participants noted how quick and straightforward it was to translate and gather data using MediBabble [Boujon et al., 2018]. In a Canadian study, MediBabble was employed as one strategy to enhance the healthcare provided for newly arrived Syrian refugees. It enabled medical doctors to examine the histories of the refugees and make diagnoses [Rahman, 2016].

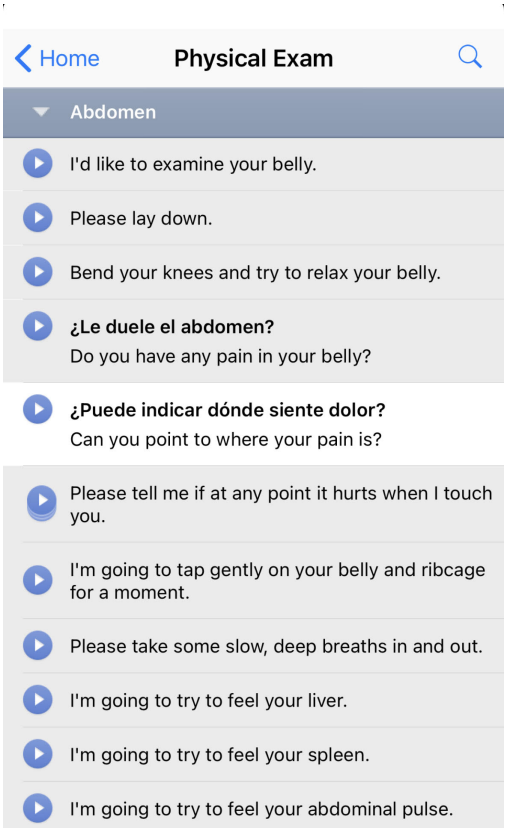

Figure 2.5: MediBabble app secondary interface

2.3.1.4 UniversalPharmacist Speaker

A variant of UniversalDoctor Speaker, UniversalPharmacist Speaker[1] has been developed to facilitate multilingual communication between pharmacists and patients and it is free to use. The app includes more than a thousand translated sentences with audio. Sentences are grouped into sub-menus, and they comprise questions and recommendations. It also offers different consecutive dialogues that facilitate the conversation between the patient and pharmacist [UniversalPharmacist, 2022]. Moreover, the web application allows to add new phrases to customize the application to the needs of pharmacists. To complete the service, UniversalPharmacist Speaker offers several pieces of written advice that can be printed in the patient's language [UniversalPharmacist, 2022] as shown in figure 2.6.

The source language is Spanish, and the app translates into German, Chinese, Flamenco, French, English, Italian, Polish, Portuguese, Romanian, Russian, Somali, Turkish, Arabic, and Moroccan Arabic. What is really interesting about the app is that it also includes questions from the patients, which pharmacists can answer immediately [UniversalPharmacist, 2022], as shown in figures 2.7 and 2.8.

[1] www.u-pharmacist.es

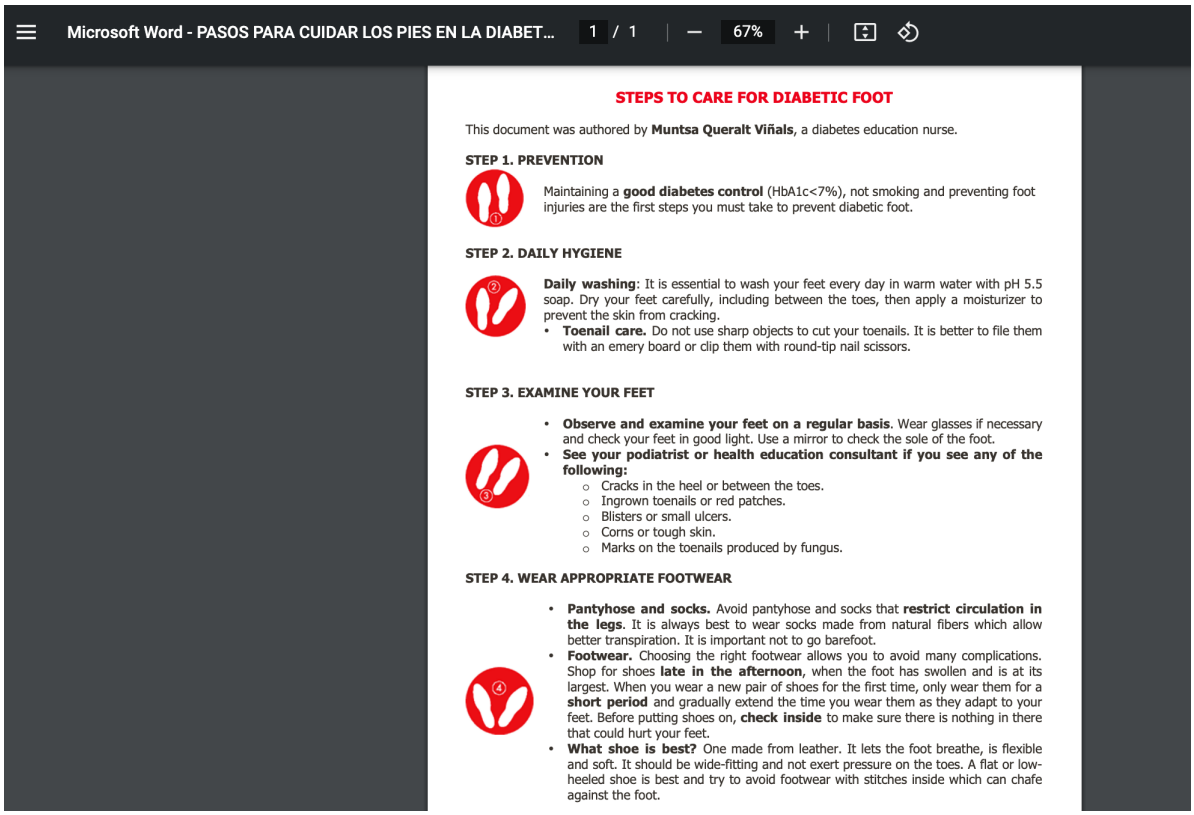

Figure 2.6: UniversalPharmacist web application screenshot: English written advice for diabetes

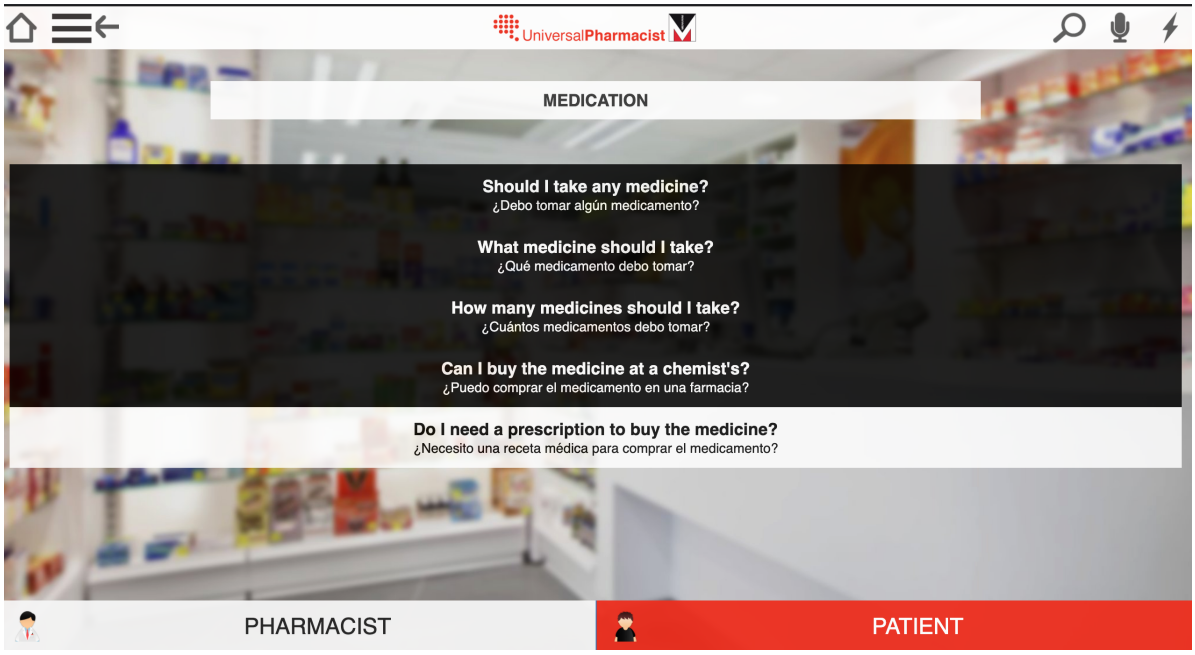

Figure 2.7: UniversalPharmacist web application screenshot: patient question

Figure 2.8: UniversalPharmacist web application screenshot: pharmacist answer

2.3.2 Machine Translation

2.3.2.1 Google Translate

Google Translate (GT) is a machine translation system that uses neural networks. A neural network is a machine learning method that predicts results from various inputs [Koehn, 2020]. Google Translate employs artificial neural networks composed of tens of thousands of separate components, or "artificial neurons," connected to tens of thousands of other artificial neurons. Each neuron in such a network responds to stimuli from other neurons and is activated following the strength, or weight, of the connections between those neurons [Kenny, 2022]. The key to successfully training a neural machine translation (NMT) system like Google Translate is to identify the exact weights that will produce the best translation model or the model whose activation states enable it to predict the most accurate translations [Kenny, 2022]. The system gains knowledge from data, just like in any machine learning. A neural translation model is gradually constructed by exposing a learning algorithm to enormous amounts of parallel data. The method adds weights over time and adjusts them so that the predictions of the model it generates are increasingly close to the desired "correct" result [Kenny, 2022].

NMT represents word meanings through plunging (= embeddings). Based on the word's distribution in the corpus or co-occurrences, the system can represent the word's meaning: the system will examine the other words that co-occur within the corpus [Koehn, 2020]. Each word is positioned in relation to the others in a multidimensional space based on how it is distributed throughout the corpus, which is why we refer to it as plunging. A string of digits corresponding to the coordinates of the word's location in space serves as its representation. Each integer represents the likelihood that the same word will appear in the context [Koehn, 2020]. This system is based on the distributional hypothesis: if two words appear in the same

context, with the same co-occurrences, they have the same meaning:

- The *tiger* climbed the tree.
- The *cat* climbed the tree.

The closer the words are to each other in the space, the more they will have the same meaning. Thanks to plunging into space, we will know which words are different and which ones will have the same meaning (an example is given in figure 2.9) [Koehn, 2020].

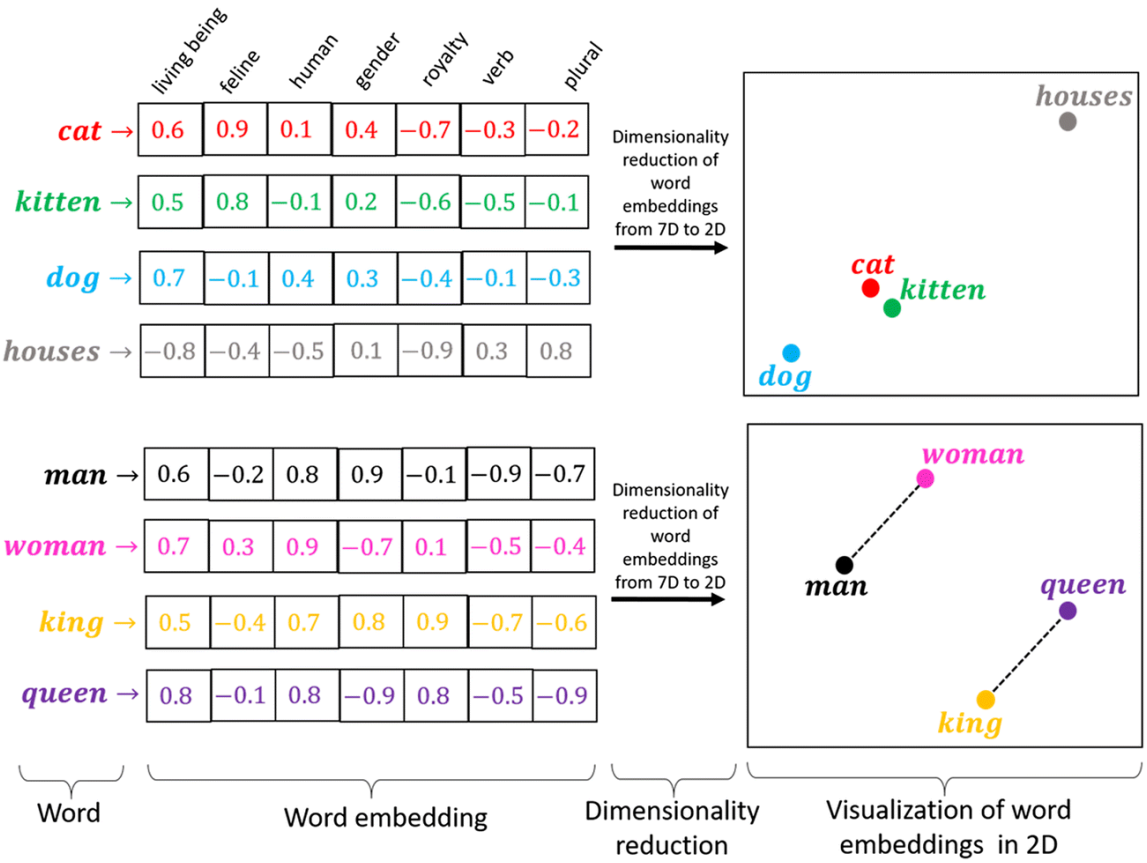

Figure 2.9: Word embedding

We can exploit similarities among languages for machine translation as we will have the same representation in all languages. If we take corpora of different languages and extract from (comparable) corpora plunges, the system puts the corresponding words in the same point of space. Nearby languages structure the information in the same way (see figure 2.10) [Koehn, 2020]. It should be mentioned that systems like Google Translate require massive amounts of bilingual text to be developed. The algorithms function even better when the amount of text supplied is large: corpora of several million words are required to achieve high-quality output [Poibeau, 2019]. The systems' performance drastically declines beyond the fifteen languages most frequently used online, especially if one of the source or target languages is not English. In these cases, the volume of data is insufficient to produce an efficient system even though NMT is a very active field of research [Poibeau, 2019]. This is related to the general lack of interest in languages other than the about fifteen widely used online, even though services like Google Translate advertise to

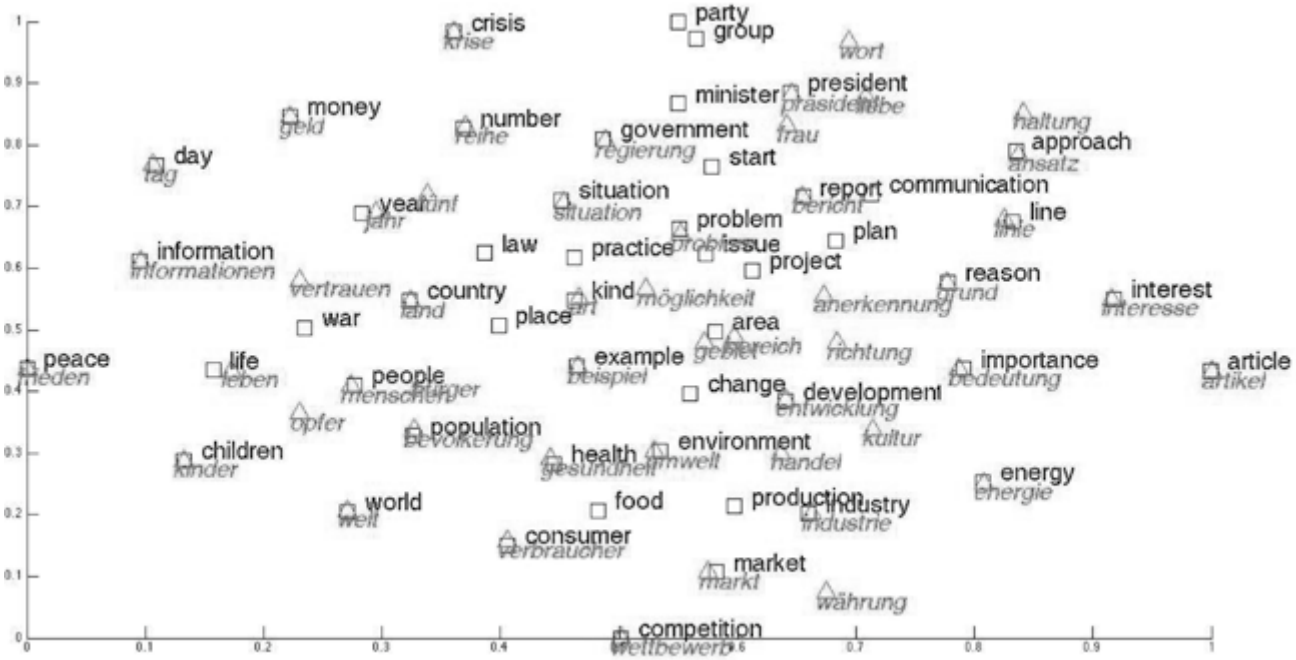

Figure 2.10: Multilingual Plunging in English and German

support more than 100 languages. The results are incredibly inconsistent and, for some of them, hardly usable [Poibeau, 2019].

One of the main advantages of NMT translation is the smoothness of the outputs that present few grammatical errors. Nevertheless, the main disadvantage is that semantic errors are frequent, as two words might always appear in the same context yet have different meanings, i.e., *cappuccino* and *espresso*. This can also result in counter-meanings and sense shifts [Poibeau, 2019]. In our case, counter-meanings represent a serious problem: in the medical domain, questions and answers should be clear of semantic errors to avoid iatrogenesis.

2.3.3 Machine Translation and Phraselators: previous studies

Many studies have been conducted comparing machine translation systems and phraselators in the medical setting, but to our knowledge, research has yet to be carried out in a pharmacy setting. Therefore, our study would be the first one to do so.

Our study follows the methodology of [Bouillon et al., 2017] who carried out a user study at Geneva University Hospitals (HUG) in 2017, where doctors were asked to use both BabelDr and Google Translate to diagnose Arabic-speaking patients. The study aimed to measure the impact of the medium (in this case, BabelDr and Google Translate) on the diagnosis. The scenarios tested were appendicitis and cholecystitis, and the language pair was French into Arabic. Although the speech recognition was good, it was found that Google Translate's translations were far less

comprehensible and less adequate than BabelDr's. Moreover, doctors expressed low confidence in Google Translate. However, doctors still reached a correct diagnosis with both systems.

[Gerlach et al., 2022] conducted a study to learn more about the impact of machine translation and speech-to-text technology on patient trust in the context of medical dialogue translation in an emergency environment. Translations of real spoken French interactions between doctor and patient were obtained from medical dialogues at the HUG and evaluated by native Albanian and Arabic speakers with refugee status. Researchers have gathered three written trustworthiness ratings for GT and BabelDr translations for each language. Results have shown that BabelDr generates a higher proportion of phrases that are totally understandable. Furthermore, BabelDr's translations were rated more trustworthy than Google Translate's [Gerlach et al., 2022].

2.4 Conclusion

In the course of this chapter we have introduced language barriers in pharmacies and in the medical context (2.1 and 2.2). We have then presented the existing applications that have been developed to overcome these languages barriers (2.3). More in detail, we have introduced different phraselators (2.3.1) such as BabelDr, CALD Assist App, MediBabble, and UniversalPharmacist Speaker. We have also presented machine translation systems and how they work with Google Translate (3.2.2). We have then concluded by summarizing two studies that have been conducted to compare BabelDr and Google Translate (2.3.3).

Chapter 3

Methodology

This chapter is structured as follows. First, we will review our different objectives and research hypotheses (section 3.1). Secondly, we will describe (i) the tests (systems configuration, language choice, and participants) (3.4) and (ii) the evaluation procedures (participants and language pairs) (3.5). Thirdly, we will focus on the quantitative and qualitative data collected (4.1 and 4.2). Lastly, we will explain the ethical considerations of this study (3.7).

3.1 Objectives and research hypotheses

This research has four objectives.

The first one aims to assess which one between BabelDr and Google Translate is the better system for triage with Arabic-speaking patients in a pharmacy setting. To do so, we will verify the following hypotheses:

- **Hypothesis 1**: Pharmacists can always reach a correct diagnosis with BabelDr and with Google Translate as previous studies have shown (2.3.3).
- **Hypothesis 2**: Pharmacists prefer employing BabelDr rather than Google Translate to perform triage as BabelDr has been specifically developed for the medical sector.

Secondly, we want to evaluate which system between BabelDr and Google Translate is the most usable in terms of successful interactions and time for performing triage in a pharmacy setting. To achieve this, we will test the following hypothesis:

- **Hypothesis 3**: BabelDr is the most usable system to perform triage in pharmacies compared to Google Translate with regards to successful interactions due to the speech recognition errors of Google Translate; with regards to time, as Google Translate is widely known and used, we hypothesize that pharmacists will reach a correct diagnosis faster with this system compared to BabelDr.

Thirdly, we aim to evaluate whether the translations of both systems are accurate, comprehensible, or dangerous for patients who only understand little standard-Arabic, but speak different vernacular-Arabic dialects such as Egyptian, Moroccan, and Tunisian. To do so, we will verify the following hypothesis:

- **Hypothesis 4**: Since both systems employ standard Arabic, only a couple of the questions will be accurate, comprehensibility scores will be low, and at least 10% of the translations will be dangerous with BabelDr and Google Translate.

Lastly, we deemed it important to have experts' opinions regarding the accuracy and dangerousness of translations with both systems: given the importance of the right medication dispensing, verifying that the systems in question employ the right terminology is essential. Hence, we will verify the last hypothesis:

- **Hypothesis 5**: From a pharmacist (domain expert) point of view, all of BabelDr's translations are accurate, and none are dangerous, but not all of Google Translate's translations are accurate, and at least 3% are dangerous.

In order to achieve our four objectives, we have set up a series of tests with pharmacists and master-level translation students to collect both quantitative and qualitative data.

3.2 Systems: settings and instructions

3.2.1 BabelDr

As previously mentioned in section 2.3.1.1, BabelDr is a flexible translator that uses speech recognition and that employs a pictograph-based response interface to interact with patients [Bouillon et al., 2021]. Pharmacists were instructed to ask simple questions and, given that the scenarios concerned different body parts, they were also instructed to change the domain on the BabelDr's application once the pain was located. This would have allowed the algorithm to perform better by offering pertinent follow-up questions [Bouillon et al., 2017].
Image 3.1 presents BabelDr's main interface: the pharmacist can decide to use the microphone to utter the question or type keywords to select the intended question from the list. Once the user has selected the chosen canonical question, a new page opens with the source question, the correspondent translation (both in written and audio form) with pictographs helping the patient answer, as shown in figure 3.2. The pharmacist can select the right pictograph and then click on the back arrow to return to the main page, where they can continue asking questions.

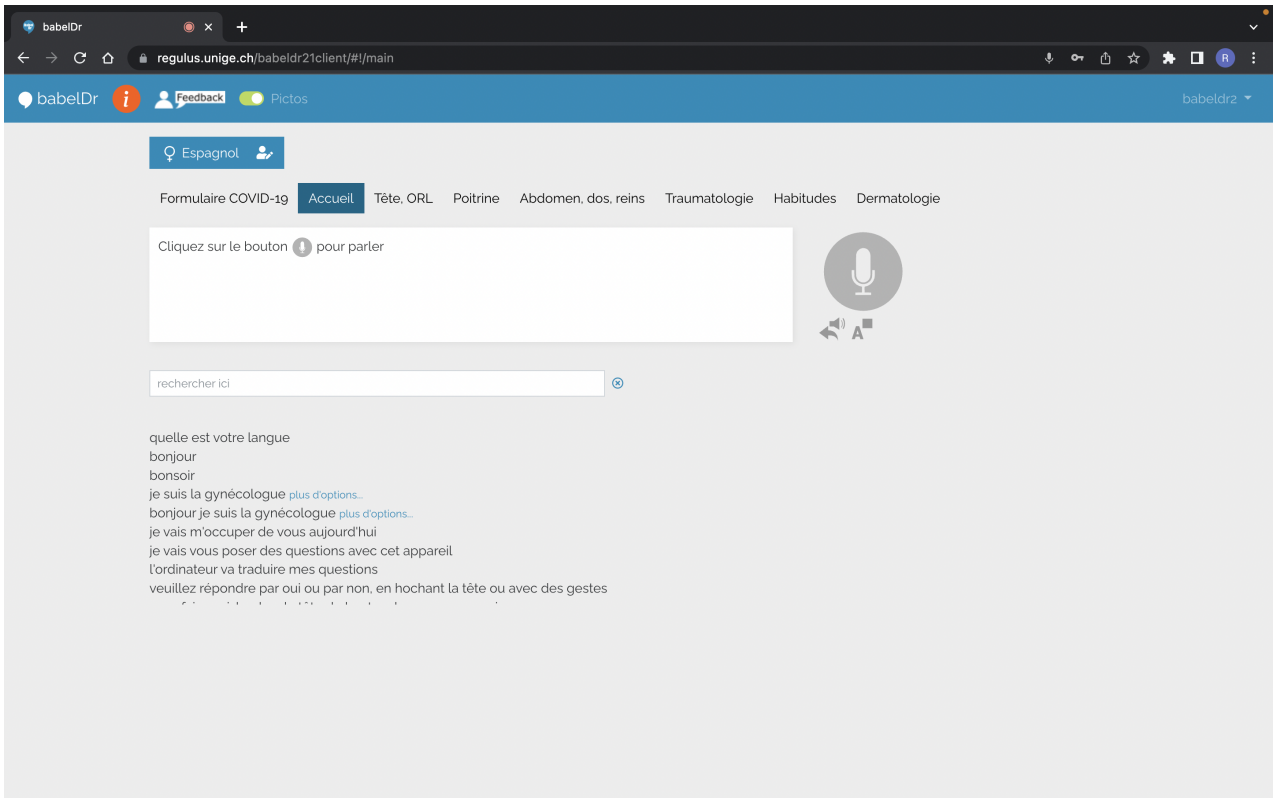

Figure 3.1: BabelDr's interface

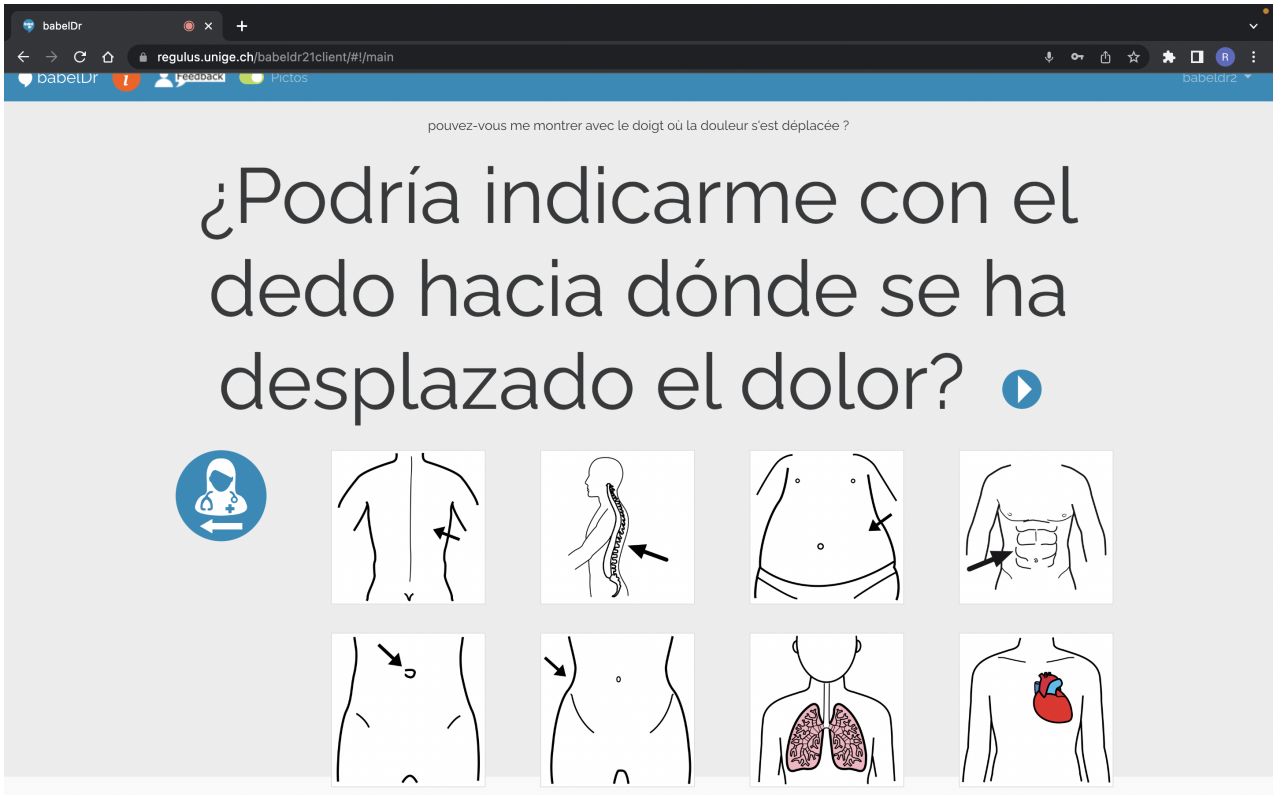

Figure 3.2: BabelDr's second interface

3.2.2 Google Translate

Google Translate's functionality was introduced in section 2.3.2.1. Compared to BabelDr, it has only one interface: as shown in figure 3.3, on the left the user can either type or record their voice and the translation appears on the right.

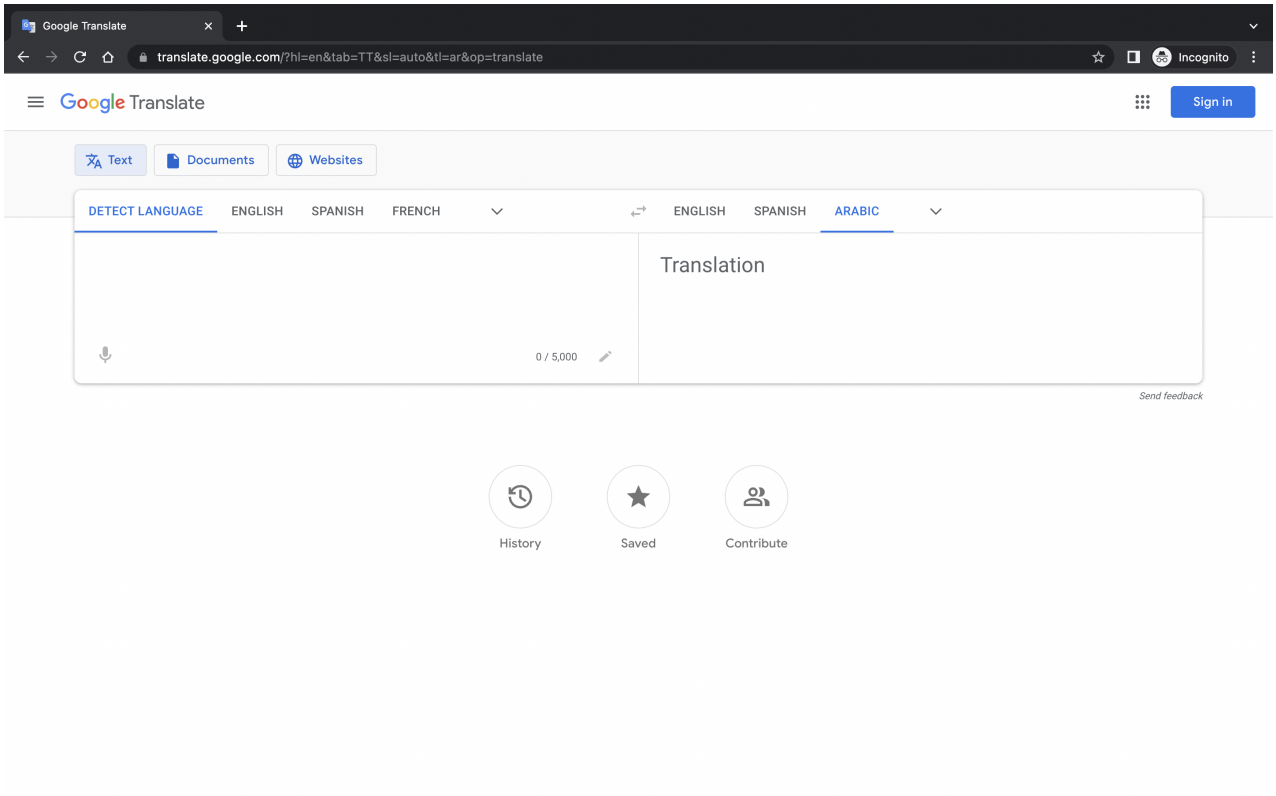

Figure 3.3: Google Translate's main interface

BabelDr and Google Translate were used on a laptop. This device was chosen as it is the environment that resembles the most the computer pharmacists would have in a real-life situation. A mouse was also provided to accurately perceive the number of clicks and how to manipulate the different systems. In order to record the tests, Zoom was employed. The *share screen* functionality was used to record the screen, the sound recording functionality to register the questions, and the internal camera to tape participants' questions and the patients' answers. It was ensured beforehand that all the questions likely to be asked were present in BabelDr.

3.3 Language choice

The chosen language pair was French-Arabic. A clarification must be made between Standard Arabic and the vernacular dialects. The ongoing entwining of religion and politics in the Arab countries is reflected in continual diglossia at the level of language and everyday communication [Kamusella, 2017]. In technical linguistic terminology, the Arabic used for writing is Modern Standard Arabic. Standard Arabic derives directly from the language of the Quran. As such, it is used for the standardization of Arabic. Throughout a millennium, this standard language has

been lost from the dialects that modern Arabic people still use in their daily lives [Kamusella, 2017]. Additionally, standard Arabic is only utilized in writing and is very infrequently spoken, suggesting that there are no native Arabic speakers. As a result, there is no standard Arabic-speaking community [Kamusella, 2017]. Standard Arabic is still a language that has been almost absent from everyday speech in Arabicphone towns and villages for fourteen centuries, both in terms of its basic grammar and lexicon. Arabs (here defined as Arabic speakers) can be found in more than 20 different vernacular speech groups, each of which is based on a particular Arabic dialect [Kamusella, 2017]. About fifty different variants are recognized by scholars and are categorized into six traditional clusters, including Mesopotamian, Levantine, Maghrebian, Egyptian, and Sudanese.

Standard Arabic is incomprehensible to an Arab who has not been educated. Nobody speaks it as a vernacular or their first language, sometimes referred to as their "native" tongue. Arabs who are educated can read and write in standard Arabic. In the same language, they speak in public and listen to university lectures. Every day, everyone hears standard Arabic recitations rooted in Quranic verses at mosque prayers or during radio or television newscasts, where the anchor reads the pertinent textual texts aloud [Kamusella, 2017]. The internet, a highly literate medium, is similar in this regard. However, among the mass media, vernaculars predominate in cinema, television, and radio. The local vernacular is enough for the poorly educated or uneducated masses to navigate daily situations. Arab women still have a high illiteracy rate, despite a significant decline from 65% in 1980 to 40% in subsequent decades [Kamusella, 2017]. Between 1980 and 2000, the illiteracy rate for women in school between the ages of 15 and 24 fell more sharply, from 45 to 19.5%. (Hammoud 2000: 20). The general illiteracy rate in Qatar, one of the wealthiest and most industrialized Persian Gulf nations, was 18.8% in 2000. Egypt, the most populous Arab nation, reportedly had a literacy rate of 45% in 2000. The overall rate of illiteracy in Arab countries was 33% in 2011 [Kamusella, 2017]. Because of this illiteracy, migrants and refugees who come to the pharmacy might not understand the Standard Arabic employed by Google Translate and BabelDr. Therefore, they might face communication problems with healthcare providers.

3.4 Tests Procedure

Cystitis, headache, and sore throat were chosen as scenarios since they are the most common in Pharma24 (internal information, 2021). For each scenario, a pharmacist and a doctor standardized a patient to ensure that all relevant symptoms and red flags were present. Each pharmacist was given two scenarios and used both systems to perform triage of the Arabic-speaking patients and eventually make a diagnosis. Participants playing the role of the patient were given a priori list of symptoms for each standardized patient and told to answer negatively to all other symptoms [Bouillon et al., 2017] (see appendix A.1).

The tests took place for five days at UniMail, in the computer lab, from the 24th of November till the 9th of December. Figure 3.4 presents the timeline of the pharmacists with the respective scenarios and the system. It was decided not to conduct pre-tests to have all the pharmacists at the same level of familiarity with

DATE	TIME	ID	SESSION 1		SESSION 2	
24-Nov	13h50 – 14h30	pharma1	*version*	BabelDr	*version*	GT
			patient	Femme	*patient*	Homme
			Cystitis		Headeache	
24-Nov	15h00 – 15h30	pharma2	*version*	GT	*version*	BabelDr
			patient	Femme	*patient*	Homme
			Headache		Sore Throat	
26-Nov	16h00 – 16h50	pharma3	*version*	BabelDr	*version*	GT
			patient	Femme	*patient*	Homme
			Headache		Sore Throat	
29-Nov	08h00 – 08h45	pharma4	*version*	GT	*version*	BD
			patient	Femme	*patient*	Homme
			Cystitis		Headeache	
29-Nov	09h00 – 09h30	pharma5	*version*	BabelDr	*version*	GT
			patient	Femme	*patient*	Femme
			Cystitis		Sore Throat	
29-Nov	16h10 – 16h55	pharma6	*version*	GT	*version*	BabelDr
			patient	Femme	*patient*	Homme
			Cystitis		Sore Throat	
08-Dec	16h00 – 16h50	pharma7	*version*	BabelDr	*version*	GT
			patient	Femme	*patient*	Femme
			Headache		Sore Throat	
09-Dec	15h30 – 16h05	pharma8	*version*	BabelDr	*version*	GT
			patient	Femme	*patient*	Homme
			Cystitis		Headeache	

Figure 3.4: Schedule and scenarios

BabelDr when carrying out the study (it was supposed that all of them had used Google Translate beforehand). Pre-tests include introducing the system by showing a demo on how to employ them and then asking the participant to test them. A maximum of one hour was foreseen with each participant. At first, participants were given a little introduction to the tests, explaining the objectives, the systems, and how these work. Next, a short tutorial followed on how to use the systems. Participants took 3-4 minutes to practice and ask questions if needed [Bouillon et al., 2017]. Later, they proceeded with testing one system with one scenario. Scenarios

and systems were balanced during the tests, each pharmacist performing a triage with Google Translate and one with BabelDr, in alternate order [Bouillon et al., 2017]. Pharmacists were given no prior information about the reasons for the visit; the only information given was that patients had pain located somewhere [Halimi and Bouillon, 2019]. This information was provided to exclude scenarios such as stress or fatigue. Pharmacists were strongly advised to ask simple yes/no questions so that patients could answer non-verbally [Spechbach et al., 2019]; the only open questions they could ask were to locate the pain, the duration of the symptoms, and the age of the patient. The experiment ended when the pharmacist reached a diagnosis and a triage decision [Halimi and Bouillon, 2019]. Subsequently they filled in a post-questionnaire. After a 2-minute break, they performed triage on the other standardized patient with the other system on another scenario and then finished by filling in the post-questionnaire. Tests lasted between 30 to 45 minutes.

3.4.1 Tests' participants

- **French speaking pharmacists**: We selected mainly pharmacists from Pharma24 Pharmacy in Geneva to conduct the tests as this pharmacy is directly adjacent to the Geneva University Hospitals and is open 24/7. Therefore it is likely to have one of the highest numbers of patients walking in every day in Geneva; in particular, this allows the pharmacist to be exposed daily to a significant variety of patients, making it more likely to encounter allophone ones. Seven pharmacists, six working at Pharma24, one at Sunstore Vernier, and one assistant working at Pharma24, were recruited. For convenience, from now on, we will always refer to these participants as "pharmacists," although note that there was also one pharmacy assistant. All work in French, although three were not native speakers, and two were perfectly bilingual (French-Arabic, French-Portuguese). All pharmacists were paid for the task.

- **Arabic-speaking patients**: two standardized Arabic-speaking patients, a female, and a male, both not speaking French, were played by four women: three university students and a mother of a university student. Participants who played the role of the patients were bilingual (AR and FR) and kindly offered to help without retribution.

The day before the experiment, participants playing the role of patients received an a priori list of symptoms they were to present, described in layman's terms. If asked questions relating to other symptoms, they were instructed to provide a noncommittal or negative answer [Spechbach et al., 2019]. The document provided to the participants playing the role of patients can be found in annex A.1.

3.4.2 Language pair for the tests

The tests' language pair was French into Standard Arabic. The reasons for testing the system for Arabic were numerous:

- Arabic was one of the languages that generated the most issues in a medical setting: refugees from Eritrea, Syria, and Afghanistan make up nearly 60% of all new asylum requests in Geneva (SEM Newsletter, October 2015) [Bouillon et al., 2017].

- Ad hoc interpreter use is more likely among speakers of Arabic, according to studies done using a sample of medical and nursing department and service leaders at Geneva University Hospitals [Bischoff and Hudelson, 2010]; in section 2.1, we have seen the implications of using ad hoc interpreters in a medical setting.

3.5 Evaluation Procedure

After the tests, we collected BabelDr's canonical sentences sent to translation and Google Translate's speech recognition results sent to translation (excluding duplicates). Finally, we asked the participants who took part in the evaluations to assess them: translation students were required to assess adequacy, dangerousness, and comprehensibility, while pharmacists were required to evaluate adequacy and dangerousness. The evaluation instructions and material were sent by email.

3.5.1 Evaluation sessions' participants

- **Translation students**: Three master-level translation students were recruited for the evaluation sessions. All study at the Faculty of Translation and Interpretation of the University of Geneva and they have standard Arabic as an active language. Furthermore, they were chosen because they come from different Arabic-speaking countries: Morocco, Egypt, and Tunisia. Hence, they know the different dialects of their respective countries. They were required to evaluate sentences based on their dialect and asses whether they would be understandable by someone from their country with very little knowledge of Standard Arabic.

- **Pharmacists**: two bilingual pharmacists took part in the evaluation session. They were recruited because they know standard Arabic, have French as their mother tongue, and have at least one year of experience in a pharmacy setting.

3.5.2 Language pair for the translation evaluations

- **Language pair for the evaluation by translation students**: in order to test the fourth hypothesis, we decided to carry out the linguistic evaluation of the translations by using three different vernacular Arabic as target languages: Tunisian, Moroccan, and Egyptian. This choice was advantageous as it allowed us to evaluate the systems in the case of patients understanding standard Arabic and in the case of patients with little understanding of the latter by only performing one set of tests. Having a system that takes into account and is understandable by vernacular-speaking Arabic patients is essential. In a Swedish study with Arabic-speaking participants, the informants expressed doubt about the interpreters' ability to convey their medical issues [Hadziabdic and Hjelm, 2014]. It was emphasized that it was crucial for the person needing an interpreter to comprehend and be understood by a competent interpreter who spoke the same language and dialect. They clarified that this was due to the major political disputes amongst the Arab nations [Hadziabdic and Hjelm, 2014]. The respondents had instances where they were provided with

a professional interpreter who was using the incorrect language or dialect, had poor translation abilities, and had not guaranteed confidentiality, which contributed to limited communication [Hadziabdic and Hjelm, 2014].

- **Language pair for the evaluation by pharmacists**: in order to test our fifth hypothesis, we asked pharmacists to evaluate the translations with the language pair French into Standard Arabic, as the translations collected with the systems were in Standard Arabic. It is important that all the technical terms and expressions are accurate and reflect the intended meaning of the pharmacists.

3.5.3 Translation quality

In order to test the last two hypothesis about the quality of the translations from the point of view of patients and pharmacists, we first need a quick review of the definition of translation quality. Koby provides a concise definition that is also broadly used in the Machine Translation (MT) field: 'A quality translation demonstrates accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account end-user needs. Thus, the main requirements can be summed up as follows: (i) fluency in the target language, which includes naturalness and grammaticality; (ii) adequacy, as in pragmatic and semantic equivalence between the source and the target text; and (iii) compliance with any requester specifications' [Koby et al., 2014]. Translation quality can be measured with automatic and human metrics.

Automated evaluations classify systems that score MT outputs without human intervention; human intervention in automated metrics occurs during task setups, such as data collection, annotations, or reference translation production. They frequently rely on reference translation-based metrics. These metrics assign a score to the MT output based on its similarity to reference translations, which are quality human translations known as gold translations [Chatzikoumi, 2020]. Compared to human evaluation, the benefits of reference translation metrics include their low cost, speed, fewer human resources, and reusability [Lavie, 2010]. Furthermore, bilingual judges are not required [Banerjee and Lavie, 2005]. Finally, the objectivity argument is widely used, to the point where this class of evaluations is referred to as 'objective evaluation' metrics as opposed to human ('subjective') evaluation metrics (Euromatrix, 2007). It is impossible to guarantee that a person will give the same score to the same text twice, nor that two or more people will agree on the evaluation. On the other hand, an automated metric will always assign the same score to the same text if all of the evaluation parameters remain constant. The consistency of automated systems is a clear advantage in this regard [Koehn, 2009]. Nonetheless, automated metrics rely on reference translations, which result from human intellect, and humans are subjective. The most common disadvantages of these metrics mentioned in the literature are the inability to distinguish between nuances [Lavie, 2010] and the difficulty in interpreting the evaluation scores [Koehn, 2009]. When reference translations are employed, the so-called reference bias occurs, where outputs from machine translation that are highly similar to the reference translation are boosted. Outputs that are not similar, even good ones, are penalized [Bentivogli

et al., 2018].

There are two subgroups within the human evaluation category: directly expressed judgment (DEJ)-based' and 'non-DEJ-based.' In DEJ-based evaluations, humans state that a translation is good, fair, or bad. In contrast, in non-DEJ-based methods, humans are asked to complete a gap-filling task based on the comprehension of a machine-translated text or to classify or correct errors of a machine-translated text [Chatzikoumi, 2020]. DEJ-based methods are subject to a greater degree of subjectivity than non-DEJ-based metrics. This is because, although there are guidelines to be followed, there is greater opportunity for subjectivity when one is asked to evaluate the quality of a text. After all, it also depends on the judge's level of (linguistic) indulgence [Chatzikoumi, 2020]. The two evaluation methods probably involve quite different cognitive processing: in DEJ-based metrics, the judges must make an assessment, but in non-DEJ-based metrics, the process is much more task-oriented (classification, postediting, question answering, and gap filling) [Chatzikoumi, 2020].

Judges, also known as annotators, are crucial in human evaluation as they have to meet specific criteria to ensure reliability. Judges can be monolingual or bilingual, native or near-native speakers of the target language or of both the source and target languages, depending on the type of evaluation. The evaluation process requires judge training, evaluation guidelines with examples, and the judge's familiarity with the field to which the texts belong [Chatzikoumi, 2020]. The following are some of the advantages of human evaluation: (i) since translations are produced for human use, human judgment is regarded as the most suitable criterion; (ii) judges can estimate the significant implications of translation errors based on their understanding of the real world; and (iii) in the case of translation, it is believed that there is no substitute for human judgment, and thus this is the standard for translation quality [Olive et al., 2011]. The most frequently mentioned disadvantage of human evaluation is its subjectivity [Olive et al., 2011]. Nonetheless, this so-called negative subjectiveness serves as the standard for the quality of automated metrics. Other drawbacks include the high cost, lack of repeatability, time-consuming nature, and low inter-annotator agreement [Olive et al., 2011].

Judges express a direct judgment on translation quality in DEJ-based evaluation methods. Accuracy (also known as adequacy in this field) and fluency are typically evaluated by comparing the source text to the target text or the target text to a reference translation [Olive et al., 2011]. When manually evaluating MT, the most commonly used methodology is to assign values from two to five-point scales representing adequacy and fluency. These scales were created for the NIST Machine Translation Evaluation Workshop, held each year [Callison-Burch et al., 2007]. While automatic measures are a valuable tool for the day-to-day development of machine translation systems, they are an inadequate substitute for human translation quality assessment [Callison-Burch et al., 2007].

For this study, which aims to help patients and pharmacists interact, we thought that human DEJ-based evaluation was the best option for evaluating the outputs of both systems. As previously said, automated evaluations would have unfairly pe-

nalized the translations if they were not similar to the gold translation even though they were correct. In our study the translations were intended to carry the same meaning as the source, whatever the form, and only human evaluators would have been able to perform such an analysis. Furthermore, given that the questions were very simple and of close-type (yes/no answers only) and related to medical symptoms, gap-filling or question-answering (typical of the non-DEJ-based evaluation) did not seem appropriate for this research. The profile of the judges was carefully selected: for the linguistic evaluation, translation master-level students were chosen as they would have provided a meticulous and attentive language-level analysis. As for the accuracy of the medical and pharmaceutical terms, two experts from the field were chosen who were both perfectly bilingual and, therefore, able to understand source and target.

In order to measure the quality of the translation output, two human evaluation tests were organized to (i) assess the adequacy, comprehensibility, and dangerousness of GT's and BabelDr's output [Halimi and Bouillon, 2019] from the point of view of the patient, and (ii) measure adequacy and dangerousness of both systems' outputs from the point of view of domain experts (pharmacists). Evaluators were asked to perform the task by using a four-point scale.

We also calculated the observed proportionate agreement for both the translators' and pharmacists' evaluations: the number of translations where annotators chose the same score, divided by the total number of translations. The more the result came closer to one, the more the judges agreed on the scores. Observed proportionate agreement was calculated for all three vernacular dialects for accuracy and comprehensibility with both systems, and for the pharmacists' evaluations it was calculated for accuracy with both systems.

When evaluating data on an interval-level scale, the Kappa coefficient is widely employed to evaluate the consistency levels of human judges [Callison-Burch et al., 2007] as it includes the probability of the agreement occurring by chance (the probability that if one judge attributes a random score, the other will say the same thing at random) as opposed to the observed proportionate agreement. The reason for measuring the Kappa score is to see if the judges agree. The more the judges agree, the more likely the evaluation is valid. The Kappa score is indispensable when reporting a human evaluation with judges, as it gives comparable and interpretable scores [Callison-Burch et al., 2007]. Cohen's Kappa is used to measure two annotators' agreement and presents a "fully-crossed design," i.e. both annotators annotate all examples. This type of Kappa was used to calculate the two pharmacists' inter-agreement. Light's Kappa is used to measure inter-agreement with more than two annotators and presents a "fully-crossed design" as well. Light's Kappa was used for the three master-level translation students [Landis and Koch, 1977].

3.6 Data collected

Two kinds of data were gathered for this research: quantitative and qualitative.

3.6.1 Quantitative data: demographic questionnaire, triage questions, diagnoses, and system usage

In order to assess which one between BabelDr and Google Translate is the better system for triage with Arabic-speaking patients in a pharmacy setting and to test the first two hypotheses connected to it, we have collected information regarding the profile of our pharmacists in order to see whether there is a correlation between them and the results of the tests.

3.6.1.1 Demographic questionnaire

The demographic questionnaire was organized into three parts: personal information, experience with technology, and experience with foreign patients. This data was useful to obtain a background of the participants to see how it affected the results and to evaluate the participants' familiarity with the technology of the study.

Personal information part comprised the following questions:

1. Sex: male; female.
2. Age
3. Mother tongue
4. Other languages
5. Hand preference
6. Years of experience
7. Profession

Experience with technology part included the following points:

1. How do you evaluate your relationship with technology? Very poor; Poor; Fair; Good; Very Good.
2. How often do you use a computer? Every day; Every 2-3 days; Once a week; More than once a month; Less than once a month; Never.
3. How would you rate your ability to use computers? Beginner; Medium; Expert.
4. Have you ever used Google Translate? Yes; No.
5. If YES, what is your opinion? Very poor; Poor; Fair; Good; Very good.
6. Have you used any other machine translation system? Yes; No; Which one?
7. Have you ever heard of voice recognition technology? Yes; No.
8. If YES, have you ever used a voice recognition system? Yes; No; Which one?

Experience with foreign patients part comprehended the following parts:

1. How often do you encounter communication problems when examining a patient with whom there is a language barrier? Never; Sometimes; Very often.

2. What strategy do you use to mitigate communication problems (you may choose more than one)? Use family and friends; Use a colleague; Use an interpreter; Other.

3. Which strategy works best (you can choose more than one)?

4. What is the most common language of patients with a language barrier (you can write more than one)?

3.6.1.2 Triage questions and diagnoses

To test the first hypothesis (pharmacist are always able to reach a correct diagnosis with BabelDr as well as with Google Translate), triage questions and diagnoses were collected and analyzed. After each session, the pharmacist communicated their diagnosis based on the responses given by patients.

3.6.1.3 System usage

To test the third hypothesis (BabelDr is more usable than Google Translate in a pharmacy setting with regards to successful interactions but not with regards to time), quantitative data about the system's usage was collected to evaluate the most efficient tool: in order to reach a diagnosis, pharmacists need to be able to interact with the system easily. Therefore, we analyzed the interactions with the system and measured the average time taken with each scenarios. With BabelDr, interactions with the system were logged as well as the text examples or recognition results the pharmacists decided to translate for the patients [Spechbach et al., 2019]. Furthermore, audio recordings for each spoken interaction and its corresponding recognition results [Spechbach et al., 2019] were collected to analyze the interactions with both systems. In particular, a special focus was given to the successful and failed interactions as these indicate how well a system performs in a pharmacy-based triage by recognizing and proposing the appropriate question formulated by the pharmacist. Finally, the duration of each session was measured by taking into account the time that had passed from the beginning of the triage until the pharmacist reached a diagnosis and triage decision. In a busy pharmacy environment, efficiency in terms of time is essential to help the most patients in the shortest time without compromising the quality of care.

3.6.2 Qualitative data: user satisfaction and translation quality evaluations

3.6.2.1 User Satisfaction

In order to assess more in detail the second hypothesis (pharmacists prefer using BabelDr instead of Google Translate to perform triage as BabelDr has been developed for the medical domain), pharmacists filled in a satisfaction questionnaire with 18 questions. These were taken from the System Usability Scale (SUS) questionnaire by John Brooke [Brooke, 1996] and adapted to BabelDr's functionalities, especially

regarding the core sentence mapping aspects and the speech recognition [Spechbach et al., 2019]. The SUS has various qualities that make its use appealing. First, there are only a few statements, making it short and simple for participants in the study to complete and for researchers to assess. Second, because it is nonproprietary, it is cheap to use and can be scored quickly after completion. Third, the SUS is technology agnostic, allowing a diverse group of usability experts to evaluate virtually any type of user interface, including those found on websites, interactive voice response (IVR) systems (both touch-tone and speech), mobile devices, and more [Bangor et al., 2009]. Bangor, Kortum, and Miller [Bangor et al., 2009] described the findings of 2,324 SUS surveys collected from 206 usability tests over ten years. In that study, it was discovered that the SUS was highly reliable and useful across a wide range of interface types. Other researchers have discovered that the SUS is a small but powerful tool for measuring usability. Tullis and Stetson [Tullis and Stetson, 2004] used five different surveys (the SUS, the Computer System Usability Questionnaire [CSUQ], the Questionnaire for User Interaction Satisfaction [QUIS], and two vendor-specific surveys) to assess the usability of two websites. They discovered that the SUS produced the most accurate results across a wide range of sample sizes.

In our tests, the questions covered: usability, learnability, appropriateness of the system to reach a diagnosis, the speech component, and the user's opinion regarding the usefulness of such a system in their daily practice [Spechbach et al., 2019]. In order to rate agreement with the questions, a five-point Likert scale (Strongly disagree, Disagree, Neutral, Agree, Strongly agree) was adopted [Spechbach et al., 2019]. The questionnaires were organized in three main parts: ease of use, performance, and personal opinion.

- **Ease of use** included this set of points:
    1. The system was easy to use.
    2. I had a hard time learning to use the system.
    3. I would expect more help from the system.
    4. I could ask my questions through the system without feeling too constrained.
    5. I liked the way the recognition result was presented.
    6. I felt comfortable interacting with the patient through the system.
    7. I could ask enough questions to be sure of my decision.
    8. Keeping track of my questions helped me in my decision.
- **Performance** had this group of points:
    1. The system reacted too slowly to my requests.
    2. The system recognized my voice easily.
    3. I trusted the translation.
    4. The system helped me reach a conclusion.
- **Personal opinion** contained these points:

1. Such a system can improve the triage process for allophone patients.
2. I could easily integrate such a system into my daily pharmaceutical practice.
3. Such a system can help pharmacists save time.

3.6.2.2 Translation quality evaluations

The evaluation corpus was composed of the eight pharmacists' questions raised during triage, disregarding the duplicates and in random order. More in detail, we selected all the 65 canonical sentences that were sent to translation for BabelDr; for Google Translate, 96 results of the speech recognition that the pharmacists sent to translation and decided to reproduce for the patient to hear. We evaluated translations from a patient's point of view and from a pharmacist point of view by using the previously-mentioned corpus.

- **Quality of translation from patients' point of view**
  To verify the fourth hypothesis of the third objective (evaluate whether the translations of both systems are accurate comprehensible, or dangerous for patients who only understand little standard-Arabic, but speak different vernacular-Arabic dialects), three Arabic master-level translation students were carefully selected for the different types of vernacular Arabic they speak: Tunisian, Moroccan, and Egyptian. They were instructed to put themselves in the shoes of someone with only a basic/little understanding of standard Arabic. This approach was beneficial mainly because migrants coming to Geneva do not necessarily have a good level of standard Arabic and as we have seen more in detail in section 3.3.

  Previous research has shown that evaluators tend to be subjective and hence come to erroneous or inconsistent conclusions if not given training or detailed instructions [Halimi and Bouillon, 2019]. Therefore, students were given a comprehensive document stating the evaluation task's aim and multiple examples to assist them (see annex A.2).

  For the first evaluation, an Excel file was offered with the following two sheets per system: *adequacy* (annex A.3 for BabelDr and annex A.5 for Google Translate) and *comprehensibility* (annex A.4 for BabelDr and annex A.6 for Google Translate).

  In the first sheet, evaluators were asked to assess the *adequacy* of the translated questions. An extract can be seen in figure 3.5. For this task, they had access to the source and the respective translation. No other reference material was provided, and the sentences were randomly ordered. They were asked to evaluate *adequacy* on a four-point scale rate: nonsense; mistranslation; ambiguous; correct [Halimi and Bouillon, 2019].

  More in detail, the scale was explained according to this criteria:

Adequacy

	Score
correct	4
ambigous	3
mistranslation	2
nonsense	1

ID	Source	Translation	Score	Dangerous
1	avez-vous des démangeaisons ?	هَلْ تَشْكِينَ مِنَ الْحَكَّة ؟		
2	la douleur est-elle comme une brûlure ?	هَلْ الأَلَمُ يُشْبِهُ الْحُرْقَة ؟		
3	prenez-vous un autre médicament ?	هَلْ تَتَنَاوَلُ دَوَاءً آخَرَ ؟		
4	avez-vous mal ailleurs ?	هَلْ تَشْعُرِينَ بِأَلَمٍ فِي أَمَاكِنَ أُخْرَى مِنَ الْجِسْمِ ؟		
5	avez-vous des allergies connues ?	هَلْ ثَمَّ تَشْخِيصُ أَيَّةِ حَسَاسِيَّةٍ لَدَيْكِ ؟		
6	avez-vous de la fièvre ?	هَلْ لَدَيْكِ الْحُمَّى ؟		
7	avez-vous des symptômes particuliers en ce moment ?	هَلْ تُلَاحِظُ ظُهُورَ أَعْرَاضٍ مُعَيَّنَةٍ الآنَ ؟		
8	pouvez-vous me montrer avec le doigt où est la douleur ?	هَلْ يُمْكِنُكِ الإِشَارَةُ بِالأُصْبُعِ إِلَى مِنْطَقَةِ الأَلَمِ ؟		
9	avez-vous une toux sèche ?	هَلْ تَشْكِينَ مِنْ سُعَالٍ جَافٍ ؟		
10	avez-vous une maladie chronique ?	هَلْ تُعَانِينَ مِنْ مَرَضٍ حَادْ ؟		
11	êtes-vous enceinte ?	هَلْ أَنْتِ حَامِل ؟		
12	la douleur est-elle intense ?	هَلْ الأَلَمُ شَدِيدٌ ؟		
13	pouvez-vous me montrer avec le doigt où est la douleur ?	هَلْ يُمْكِنُكِ الإِشَارَةُ بِالأُصْبُعِ إِلَى مِنْطَقَةِ الأَلَمِ ؟		
14	depuis combien de jours ?	مُنْذُ كَمْ يَوْمٍ ؟		
15	bonjour	مَرْحَبًا		
16	avez-vous mal quand vous urinez ?	هَلْ تَشْعُرِينَ بِأَلَمٍ عِنْدَ التَّبَوُّلِ ؟		
17	quelle est l'intensité de la douleur sur une échelle de zéro à dix, zéro étant le minimum et dix le maximum ?	مَاهِيَ شِدَّةُ الأَلَمِ عَلَى مِقْيَاسِ دَرَجَاتٍ مِنْ 0 إِلَى 10 عِلْمًا أَنَّ 0 يَعْنِي غِيَابَ الأَلَمِ وَ 10 الْحَدَّ الأَقْصَى لِلأَلَمِ ؟		
18	avez-vous pris un médicament ?	هَلْ تَنَاوَلْتِ دَوَاءً مَا ؟		

Figure 3.5: Extract of the adequacy evaluation sheet

- correct: the sentence has the same meaning in the source and the target and is well constructed;
- ambiguous: the sentence is correct but can be misinterpreted;
- mistranslation: the translation does not have the same meaning as the source, but it is still syntactically and lexically correct;
- nonsense: the meaning is not the same and and the sentence is syntactically and lexically incorrect;

In particular, they were asked to rate the *adequacy* of the translation by indicating if the meaning of the original text matched the translation [Halimi and Bouillon, 2019] and to consider if the translations were culturally adequate and would therefore make sense to someone who had just a little understanding of the standard Arabic. Evaluators were clearly instructed not to evaluate whether the translations matched the source question perfectly but rather to check if the meaning was the same, i.e., if the source sentence was "do you have fever?" and the translation had been "are you feverish?", the translation would have to be considered correct since the two sentences are semantically equal. In the same sheet, they were also asked to assess the *dangerousness*, namely the translations that have an opposite meaning from the source sentence, which could have an impact on the pharmacist's diagnosis since they lead to answers that do not match the intended objective [Halimi and Bouil-

lon, 2019].

Evaluators also assessed the *comprehensibility* in the second sheet. They were only given the Arabic translations for this task since the French reference might have influenced their evaluation. The task evaluated the ease with which patients would understand the question [Halimi and Bouillon, 2019]. Again, they were given a four-point scale rate: incomprehensible, syntax error, non-idiomatic, fluent. An extract of this sheet can be seen in figure 3.6. In a medical setting, comprehensibility plays a key role as it measures to what extent the information has been received without considering if the translation has the correct meaning [Halimi and Bouillon, 2019].

Comprehensibility

	Score
fluent	4
non-idiomatic	3
syntax error	2
incomprehensible	1

ID	Translation	Score
1	هَلْ تَشْكِينَ مِنَ الْحَكَّة ؟	
2	هَلْ الأَلَمُ يُشْبِهُ الْحُرْقَة ؟	
3	هَلْ تَتَنَاوَلُ دَوَاءً آخَرَ ؟	
4	هَلْ تَشْعُرِينَ بِأَلَمٍ فِي أَمَاكِنَ أُخْرَى مِنَ الْجِسْمِ ؟	
5	هَلْ تَمَّ تَشْخِيصُ أَيَّةِ حَسَاسِيَّةٍ لَدَيْكَ ؟	
6	هَلْ لَدَيْكَ الْحُمَّى ؟	
7	هَلْ تُلَاحِظُ ظُهُورَ أَعْرَاضٍ مُعَيَّنَةٍ الآنَ ؟	
8	هَلْ يُمْكِنُكِ الإِشَارَةُ بِالأُصْبُعِ إِلَى مِنْطَقَةِ الأَلَمِ ؟	
9	هَلْ تَشْكِينَ مِنْ سُعَالٍ جَافٍ ؟	
10	هَلْ تُعَانِينَ مِنْ مَرَضٍ حَادْ ؟	
11	هَلْ أَنْتِ حَامِل ؟	
12	هَلْ الأَلَمُ شَدِيدٌ ؟	
13	مُنْذُ كَمْ يَوْمٍ ؟	
14	مَرْحَبًا	
15	هَلْ تَشْعُرِينَ بِأَلَمٍ عِنْدَ التَّبَوُّلِ ؟	
16	مَاهِيَ شِدَّةُ الأَلَمِ عَلَى مِقْيَاسِ دَرَجَاتٍ مِنْ 0 إِلَى 10 عِلْمًا أَنَّ 0 يَعْنِي غِيَابَ الأَلَمِ وَ10 الْحَدَّ الأَقْصَى لِلْأَلَمِ ؟	
17	هَلْ تَنَاوَلْتِ دَوَاءً مَا ؟	
18	هَلْ تَعَرَّضْتَ لِضَرْبَةٍ فِي الرَّأْسِ ؟	
19	مُنْذُ كَمْ يَوْمٍ تَشْعُرُ بِالأَلَمِ ؟	

Figure 3.6: Extract of the comprehensibility evaluation sheet

- **Quality of translation from pharmacists' point of view**
  We organized a second evaluation with two pharmacists (experts in this research domain) who did not take part in the tests to verify the fifth hypothesis: from a pharmacist point of view, all of BabelDr's translations are accurate and none are dangerous, but not all of Google Translate's translation are accurate and at least 3% are dangerous. Both experts were bilingual and came from different Arabic-speaking countries: Tunisia and Syria. This evaluation helped assess the point of view of an expert who bases their diagnosis on the patient's answers and has a deeper understanding of the technical terms. The fact that the two experts come from different countries with different dialects should not have affected the results, given that they have a perfect level of Arabic and were asked to evaluate the translations by taking standard Arabic into account.

  Pharmacists were given the same excel file as the translators, asking to evaluate the *adequacy* (the extract is the one in figure 3.5). They had access to the source and the corresponding translations, and the scale rate was the same as in the first evaluation with the translation students. They were also asked to indicate the *dangerousness* of the translations. By not being linguistic experts, it was not essential to have their judgment on fluency.

3.7 Ethical considerations

The Cantonal Research Ethics Commission CCER, Health Department, and Cantonal Pharmacist's Office were consulted to seek ethical approval. In their opinion, this project does not fall within the scope of the law on research on human beings (LRH) and does not need to be submitted to the commission. Indeed, this project does not aim at obtaining generalizable knowledge on human diseases or the structure and functioning of the human body. The non-entry into the matter implies that the CCER does not pronounce itself on the progress of this project. However, in their opinion, everything indicates that the study would follow the general ethical principles applicable to all research involving people.

Participation in the study was voluntary, and pharmacists were remunerated. Moreover, all collected data has been coded and will be deleted after a year from their collection. The participant's results will remain confidential, and there will be no possibility of associating or identifying the participants with the results of this study in this work or future publications.

3.8 Conclusion

In this chapter, we have outlined our methodology. We began by defining our research objectives and hypotheses (section 3.1). Secondly, we described the tests by presenting the systems configuration, the language pair choice, and participants (3.4); we then continued illustrating the evaluation participants and language pairs, with a small focus on translation quality (3.5). Thirdly, we presented the quantitative and qualitative data collected (4.1 and 4.2). Lastly, we explained the ethical

considerations (3.7).

For the reader's convenience, Table 3.1 summarizes our research hypotheses and the data we used to test them.

Research hypothesis	**Data collected**	**Type of data**
Pharmacists can always reach a correct diagnosis with BabelDr as well as with Google Translate	Triage questions, diagnoses	Quantitative data
Pharmacists prefer employing BabelDr rather than Google Translate to perform triage as BabelDr has been specifically developed for the medical sector	User satisfaction	Qualitative data
BabelDr is the most usable system to perform triage in pharmacies compared to Google Translate with regards to successful interactions but not with regards to time	System usage	Quantitative data
Since both systems employ standard Arabic, only a couple of the questions will be accurate, comprehensibility scores will be low, and at least 10% of the translations will be dangerous with BabelDr and Google Translate	Translation quality	Qualitative data
From a pharmacist (domain expert) point of view, all of BabelDr's translations are accurate, and none are dangerous, but not all of Google Translate's translations are accurate, and at least 3% are dangerous	Translation quality	Qualitative data

Table 3.1: Research hypothesis and type of data collected

Chapter 4

Results

This chapter will present the results of our experience. First, we will introduce the quantitative data collected (section 4.1), starting with the demographic questionnaire (4.1.1), diagnosis and triage questions (4.1.2), and system usage (4.1.3). The latter will be divided into two subsections: interactions with the system (4.1.3.1) and time (4.1.3.2). Secondly, it will present the collected qualitative data: results of the user satisfaction (4.2.1) and the translation quality results of the translation students (4.2.2) and the experts (4.2.3).

4.1 Quantitative data

4.1.1 Demographic questionnaire

This section will focus on the results of the demographic questionnaire presented in section 3.6.1.1. This data was collected in order to present the profile of the pharmacists who participated in the tests and to assess whether there is any correlation between their profile and the results.

4.1.1.1 Personal information

1. Six female and two male pharmacists participated in the tests;
2. Two of the participants were between 18-25 years old, and the other six were aged between 26-35 years old;
3. Seven participants had French as a native language, two had also Arabic, another two had also Portuguese, and one had Italian;
4. Participants declared being able to speak other languages such as English, French, Spanish, and German.
5. Six participants were right-handed, and two were left-handed;
6. Years of experience at the pharmacy varied from 7 years to 1 month.
7. All but one of the participants were pharmacists, the one being a pharmacy assistant.

4.1.1.2 Experience with technology

1. Half of the participants evaluated their relationship with technology as good, and the other half as very good;
2. All the participants affirmed that they use a computer daily;
3. Five participants affirmed to be medium experts and three to be experts with computers;
4. All of them have already used and know Google Translate;
5. Regarding their opinion of Google Translate, half of them reckoned that it is fair, and the other half rated it as good;
6. All of the participants declared having heard of voice recognition technology;
7. Four participants affirmed having already used a voice recognition system like Siri, Google, and Google Translate.

4.1.1.3 Experience with foreign patients

1. Three pharmacists affirmed encountering language barriers sometimes, three often, and two very often;
2. The most common strategy used to mitigate communication problems was asking a colleague for help, and the second most common was asking the patient's family or friends; two pharmacists also affirmed having used Google Translate, and one said to have employed DeepL.
3. The strategy that is considered to work the best is asking a colleague but also having the help of family and friends; the other strategies mentioned were calling a phone translation service;

4.1.2 Diagnosis and triage questions

Our hypothesis stated that pharmacists always reach a correct diagnosis with BabelDr as well as with Google Translate. The results of our research confirm this hypothesis: pharmacists reached a correct diagnosis in all 16 sessions based on the information they collected with both systems. This result shows how both systems suited the task and allowed pharmacists to collect data correctly. The following section will present each scenario's most frequently asked questions. Please note that only questions or core sentences asked more than once will be listed as they are the most relevant and most likely to be asked in the future. This might be important to future systems' developments.

4.1.2.1 Scenario 1: Cystitis

With BabelDr, the cystitis scenario required, on average, 16 questions, considering all the oralized questions and not just the canonical forms sent to translation. In contrast, Google Translate required, on average, 15 questions. With both systems we counted also the rejected canonical forms (BabelDr) and the speech recognition

results not sent to translation (GT). For the first scenario, we can affirm that the system did not influence the number of questions asked.
Three pharmacists tested the cystitis scenario with BabelDr. Table 4.1 will present the most frequently canonical questions sent to translation sorted by frequency. The two questions more frequently asked with BabelDr were "does it hurt when you urinate?" and "for how many days?". Two pharmacists have tested the cystitis scenario with Google Translate. Table 4.2 shows the most frequently asked questions for this scenario sorted by frequency with the latter system. The list, in this case, is shorter since every pharmacist formulated the questions differently, and it was rare having two questions formulated the same way. For this system, the two most commonly asked questions were "where does it hurt?" and "how long has it been?".

Original French	**English translation**
avez-vous mal quand vous urinez?	does it hurt when you urinate?
depuis combien de jours?	for how many days?
pouvez-vous me montrer avec le doigt où est la douleur?	can you show me with your finger where the pain is?
avez-vous pris un médicament?	did you take a medication?
êtes-vous enceinte?	are you pregnant?
la douleur est-elle comme une brûlure?	is the pain like a burn?
c'est la première fois que ça vous arrive?	is this the first time this has happened to you?
y a-t-il du sang dans les urines?	is there blood in the urine?
avez-vous mal dans le bas du dos?	do you have lower back pain?

Table 4.1: Cystitis: most frequent questions with BabelDr.

Original French	**English translation**
où est-ce que vous avez mal?	where does it hurt?
depuis combien de temps?	how long has it been?
vous avez déjà pris des médicaments?	have you already taken some medication?
est-ce que vous avez de la fièvre?	do you have a fever?

Table 4.2: Cystitis: most frequent questions with Google Translate.

Recordings of this scenario show that locating the pain was the difficult part: in fact, Muslim women avoided to indicate precisely their genitalia so pharmacists tended to ask multiple times where the area was with both systems. There seems to be no pictograph of women genitalia in BabelDr's system.

4.1.2.2 Scenario 2: Headache

Three pharmacists tested the headache scenario with both systems. On average, with BabelDr, this scenario required 17 oralized questions, whereas on average 15 questions were oralized with Google Translate. Table 4.3 illustrates the most frequently canonical questions sent to translation with BabelDr sorted by frequency, whereas table 4.4 illustrates the most common questions with Google Translate. The two most common questions asked with BabelDr were "can you show me with your finger where the pain is?" and "for how many days?". The two most common with Google Translate were "have you already taken some medication?" and "for how many days?".

Original French	**English translation**
pouvez-vous me montrer avec le doigt où est la douleur?	can you show me with your finger where the pain is?
depuis combien de jours?	for how many days?
avez-vous reçu un coup à la tête?	did you receive a blow to the head?
avez-vous d'autres plaintes?	do you have any other complaints?

Table 4.3: Headache: most frequent questions with BabelDr.

Original French	**English translation**
est-ce que vous avez déjà pris un médicament pour le mal de tête?	have you already taken a medication for the headache?
depuis combien de jours?	for how many days?
est-ce que vous avez mal autre part?	do you have pain anywhere else?
est-ce que vous avez d'autres symptômes?	do you have any other symptoms?
est-ce que c'est la première fois que ça vous arrive?	is this the first time this has happened to you?

Table 4.4: Headache: most frequent questions with Google Translate.

4.1.2.3 Scenario 3: Sore throat

Two pharmacists tested the third scenario with BabelDr with an average of 13 oralized questions, whereas three pharmacists tested it with Google Translate with an average of 14 oralized questions. Tables 4.5 and 4.6 present the most frequently canonical questions sent to translation sorted by frequency with BabelDr and Google Translate, respectively. With BabelDr, the two most common questions were "can you how me with your finger where the pain is?" and "for how many days?", whereas with Google Translate they were "do you have a fever?" and "do you find it hard to swallow?".

Original French	**English translation**
pouvez-vous me montrer avec le doigt où est la douleur?	can you show me with your finger where the pain is?
depuis combien de jours?	for how many days?
avez-vous de la fièvre?	do you have a fever?

Table 4.5: Sore throat: most frequent questions with BabelDr.

Original French	**English translation**
est-ce que vous avez de la fièvre?	do you have a fever?
est-ce que vous avez du mal à avaler?	do you find it hard to swallow?
où avez-vous mal?	where does it hurt?
depuis combien de jours?	for how many days?
avez-vous une maladie chronique?	do you have a chronic disease?

Table 4.6: Sore throat: most frequent questions with Google Translate.

4.1.3 System Usage

Our third hypothesis stated that BabelDr is the most usable system to perform triage in pharmacies compared to Google Translate with regards to successful interactions but not with regards to time. Our research disproved this hypothesis: Google Translate proved to be the most efficient system in terms of successful interactions (results are presented more in detail in section 4.1.3.1). As far as time is concerned, for each scenarios both systems performed almost the same.

4.1.3.1 Interactions with the system

Table 4.7 presents the interactions with both systems. The number of interactions was mostly the same: 135 oralized questions with BabelDr and 139 oralized questions with Google Translate. Table 4.7 shows that 93% of oralized questions were accepted and sent to translation on Google Translate against the 87% of utterances sent to translation on BabelDr. Patients answered to all the questions sent to translation.

It is complex to define successful interaction since pharmacists do not understand the target language (in this case, Arabic), and the two systems function differently: pharmacists can only judge the correctness of the speech recognition and how many times they had to ask the same question [Bouillon et al., 2017]. Hence, in this study, a successful interaction will be considered when the pharmacist approved the speech recognition either by playing the translation in Google Translate or by sending the canonical sentence to translation in BabelDr. However, this does not entail that the recognition matches perfectly the spoken utterance, but rather that the user considered that the recognition matched the intended meaning [Bouillon et al., 2017].

System	Oralized questions	Accepted questions sent to translation
BabelDr	135	117 (87%)
GT	139	130 (93%)

Table 4.7: Number of interactions with the system.

If we analyze more in detail the interactions that were not successful with BabelDr, we can see that the causes are different. Table 4.8 will help us illustrate these interactions.

The first group presents the utterances that were not oralized because they were out of coverage (1) [Bouillon et al., 2017]. We can find two types of these. On the one hand, there are the interactions that were not present in the domain coverage of the system among the canonical sentences included (1a) [Bouillon et al., 2017]. These included mostly vague questions such as "Have you tried anything?". On the other hand, we can find interactions that presented surface forms that were not covered by the grammar (1b) [Bouillon et al., 2017]. The latter were either due to gaps in the system coverage or to pharmacists using informal language (such as "Are you in much pain or are you okay?") or complex sentences which resulted in incorrect recognition results.

Failed interactions for in-coverage utterances are included in the second group (2) [Bouillon et al., 2017]. Some utterances were rejected because pharmacists did not find the canonical pertinent (2a) [Bouillon et al., 2017]. The other failed interactions (2b) were due to speech recognition errors caused by a long silence at the beginning of the interaction [Bouillon et al., 2017].

Finally, interactions failed because of bad recording can be found in the third group (3).

1. Out of coverage	
a. Out of domain	5
b. Out of grammar	4
2. In coverage	
a. Canonical rejected	5
b. Recognition error	3
3. Interaction issues	**1**
Total non translated	18

Table 4.8: Non-oralized translations with BabelDr

With Google Translate, failed interactions were considered those that were recognized by the system but were not played by the pharmacist for the patient to hear. They were always due to speech recognition errors. In particular, homophones proved to be a challenge for Google Translate: in the question "Avez-vous de la toux?" (Do you have a cough?) "toux" was misrecognized as "tour" (tower) or "tout" (everything), or "Est-ce que vous alaitez?" (Are you breastfeeding?), recognized as "Est-ce que vous allez tu es?" (Are you going to be?). The reason might be that Google Translate was not explicitly developed for the medical domain. Therefore it proposes the most probable words based on the corpora it has been trained on (as explained in section 3.2).

Pharmacists did not send any of the failed interactions to translation, and they rather reformulated the question.

There does not seem to be any correlation between the native language of the participants and the failed interactions.

4.1.3.2 Time

Table 4.9 shows the average time to complete each scenario. According to our third hypothesis, Google Translate should have been the most usable system in terms of time as this system is widely used and all participants declared to have already used Google Translate in the demographic questionnaire.

It must be pointed out that the time does not necessarily reflect the system's performance, but it could also be interpreted as each pharmacist's personal and professional choice as to how much time they decided to take to ask for all the red flags. For example, one pharmacist spent 660 seconds for the second scenario with BabelDr, and 960 seconds with Google Translate for the third scenario. In contrast, another pharmacist with the same scenarios spent 420 seconds with BabelDr and

System	Cystitis	Headache	Sore throat
BabelDr	300 sec	480 sec	420 sec
GT	300 sec	420 sec	480 sec

Table 4.9: Average time to complete the scenarios.

300 seconds with Google Translate. Nevertheless, it is interesting to point out that although BabelDr needs a bit more manipulation on the user's part (activating the microphone, selecting the canonical question, selecting a pictograph, coming back), this did not negatively influence the time of the triage.

4.2 Qualitative data

4.2.1 User Satisfaction

Our second hypothesis stated that pharmacists prefer using BabelDr for triage in pharmacy rather than Google Translate as BabelDr has been specifically developed for the medical domain. The results of the User Satisfaction proved our hypothesis to be true in some aspects: BabelDr is the preferred system as far as performance and personal opinion is concerned. However, if we consider ease of use, both systems proved to be appealing at the same level. In the analysis of the answers below, for the sake of simplicity and clarity, we will only focus on the "strongly agree" and "strongly disagree" answers. Nevertheless, please note that results have been more nuanced than this. For more detail, the figures of this section offer more information about the pharmacists' answers.

4.2.1.1 Ease of use

Figures 4.1 and 4.2 present the results concerning the ease of use.

1. **The system was easy to use**: Four participants strongly agreed with the affirmation both for Google Translate and BabelDr, however one participant disagreed with the affirmation for BabelDr;

2. **I had a hard time learning to use the system**: Google Translate proved slightly more intuitive, with five pharmacists strongly disagreeing with the affirmation, while four strongly disagreeing for BabelDr;

3. **I would expect more help from the system**: in this case there is no strong opinion, however two participants agreed with the fact that they would have expected BabelDr to help more and none agreed for Google Translate;

4. **I could ask my questions through the system without feeling too constrained**: for Google Translate, two participants strongly agreed with the affirmation, while only one strongly agreed for BabelDr;

5. **I liked the way the recognition result was presented**: five participants strongly agreed with the affirmation for BabelDr, but only one strongly agreed for Google Translate.

Figure 4.1: BabelDr US: Ease of Use

6. **I felt comfortable interacting with the patient through the system**: No strong opinion for both systems in this case, however, four participants agreed with the assessment for BabelDr, and three agreed for Google Translate.

7. **I could ask enough questions to be sure of my decision**: BabelDr in this case was the preferred system: three participants strongly agreed; whereas for Google Translate, two participants strongly agreed.

Figure 4.2: Google Translate US: Ease of Use

8. **Keeping track of my questions helped me in my decision**: Keeping track of the questions was seen as more efficient with BabelDr (four strongly agreed), while this was not the case for Google Translate (one strongly agreeing and one strongly disagreeing).

4.2.1.2 Performance

Figures 4.3 and 4.4 present the results concerning the performance of both systems.

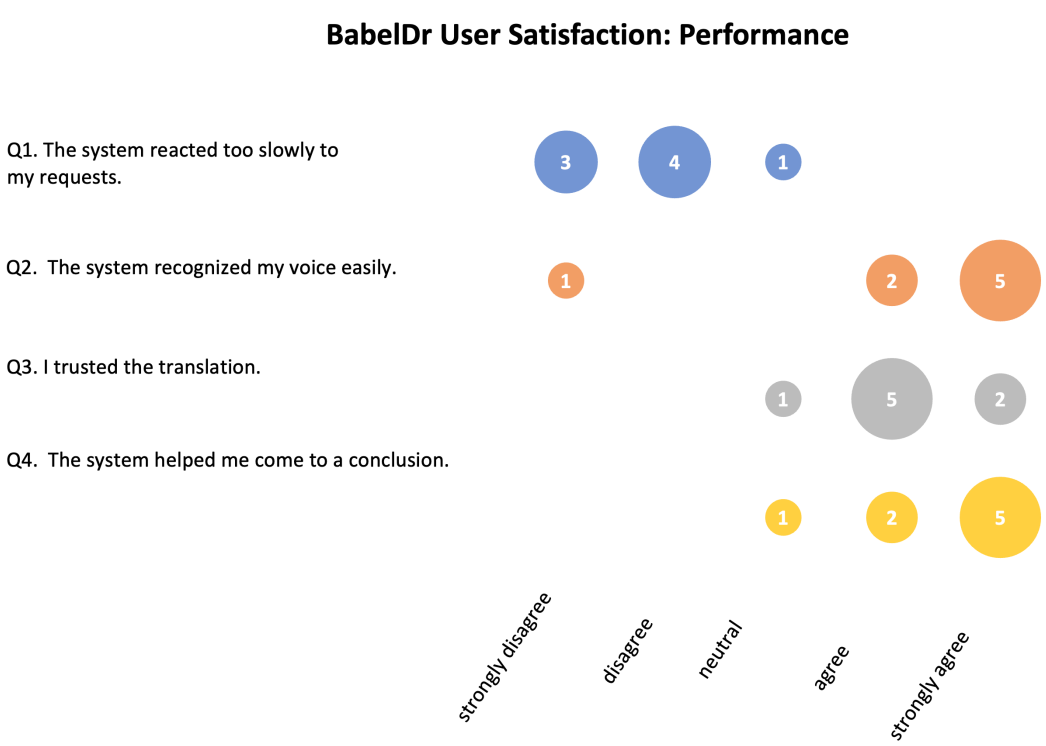

Figure 4.3: BabelDr US: Performance

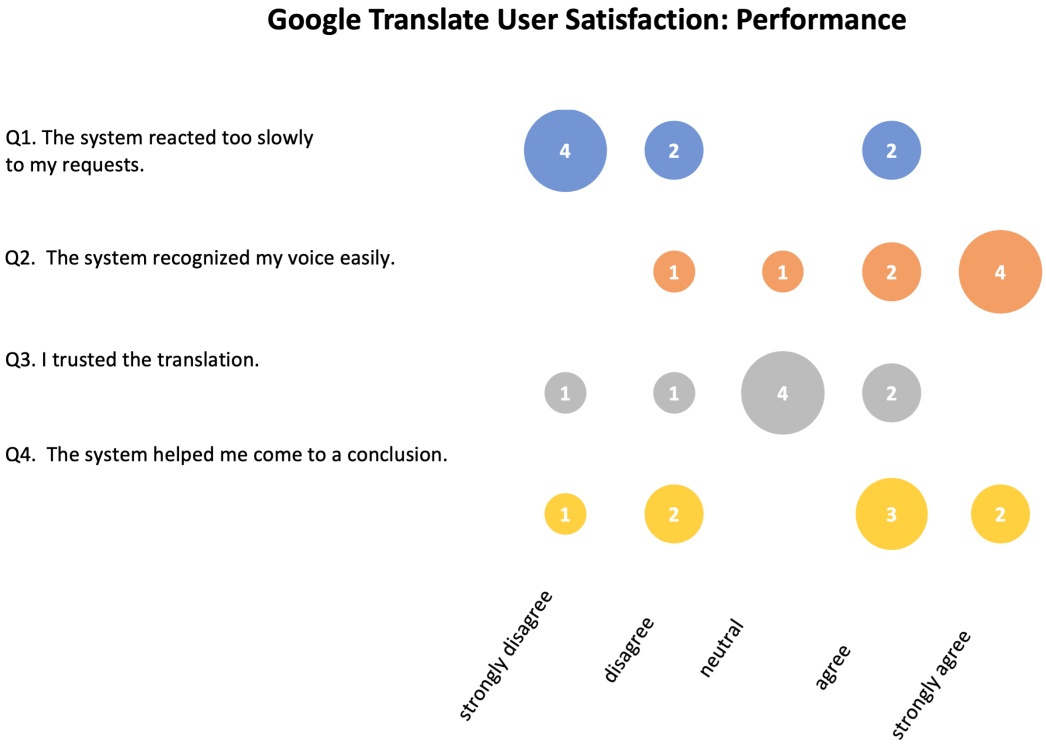

Figure 4.4: Google Translate US: Performance

1. **The system reacted too slowly to my requests**: We can see how partic-

ipants thought that GT was slightly faster in the reaction, with four people strongly disagreeing with the affirmation for GT while three strongly disagreeing for BabelDr.

2. **The system recognized my voice easily**: We can see how pharmacists preferred the voice recognition of BabelDr (five strongly agree) to the one of GT (four strongly agree).

3. **I trusted the translation**: It is evident that participants trusted the translations of BabelDr (two strongly agree) than those of Google Translate (zero strongly agreeing and one strongly disagreeing); this might be since, in the introduction, it was pointed out that professional translators and interpreters did the translations in BabelDr.

4. **The system helped me reach a conclusion**: BabelDr turned out to be the preferred system in this case with five participants strongly agreeing with the affirmation for BabelDr, while two strongly agreeing for Google Translate.

4.2.1.3 Personal Opinion

Figures 4.5 and 4.6 present the results concerning the personal opinion for both systems.

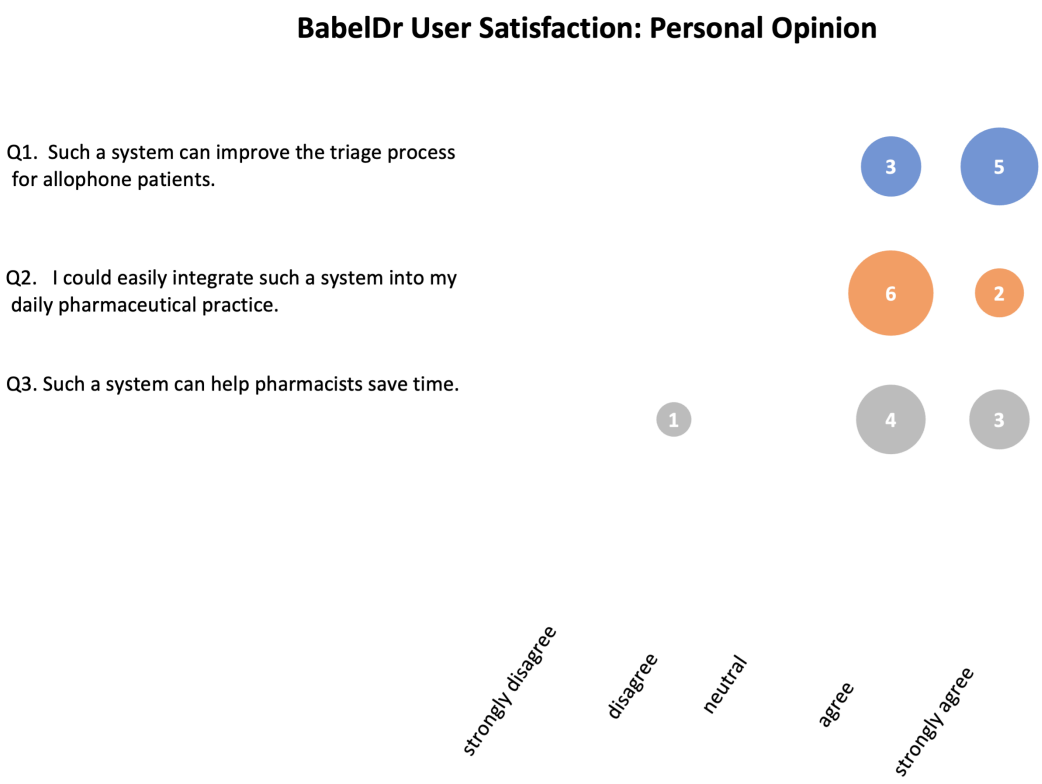

Figure 4.5: BabelDr US: Personal Opinion

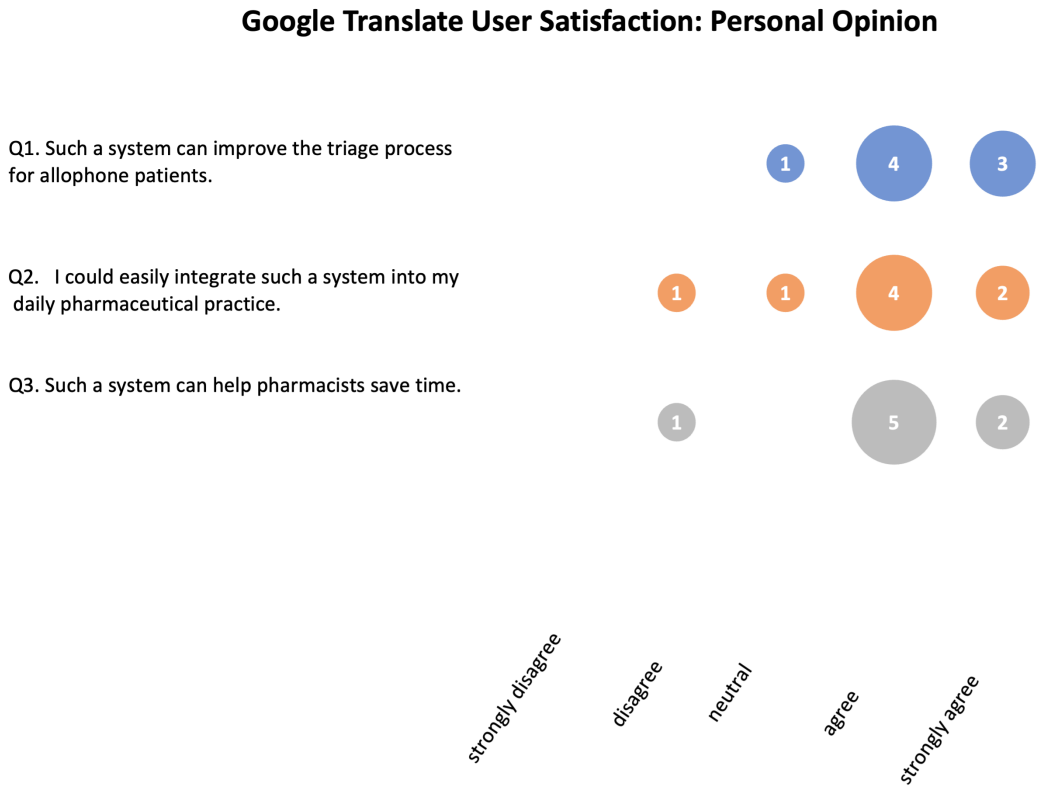

Figure 4.6: Google Translate US: Personal Opinion

1. **Such a system can improve the triage process for allophone patients**: five participants strongly agreed with the assessment for BabelDr, while three strongly agreed for Google Translate.

2. **I could easily integrate such a system into my daily pharmaceutical practice**: for both systems, two participants strongly agreed with the affirmation, however one disagreed for Google Translate.

3. **Such a system can help pharmacists save time**: three participants strongly agreed with the assessment for BabelDr and two strongly agreed for Google Translate.

Table 4.10 summarizes the results of the User Satisfaction.

As far as ease of use is concerned, the hypothesis that pharmacists prefer BabelDr to Google Translate is confuted: participants seem to have no absolute preferred system. Out of 8 criteria, BabelDr proved to be the preferred system for half of them. As far as performance is concerned, out of four criteria, BabelDr proved to be the preferred system for three. In fact, Google Translate was thought to react faster to the requests. As far as personal opinion is concerned, pharmacists preferred BabelDr with all three criteria.

	BabelDr	**Google Translate**
Ease of use		
Easiest system to use		x
Easiest system to learn to use the system		x
The system that helped the most		x
The system that allowed to ask questions without feeling too constrained		x
The system that presented better the recognition result	x	
The preferred system to interact with the patient	x	
The system that allowed to ask enough questions to be sure of the decision	x	
The system that helped keeping track of the questions	x	
Performance		
The system that reacted the fastest to the requests		x
The system that recognized the voiced more easily	x	
The system that offered a trusted translation	x	
The system that helped reaching a conclusion	x	
Personal opinion		
The system that can improve the triage process for allophone patients	x	
The system that can be easily integrated in the daily pharmaceutical practice	x	
The system can help pharmacists save time	x	
TOTAL	**10**	**5**

Table 4.10: Summary of the user satisfaction: the cross next to the affirmation indicates the system that has received the highest number of "strongly agree/strongly disagree", and in case there are none, the highest number of "agree/disagree".

4.2.2 Translators' Evaluations

By collecting the translator's evaluations as described in section 3.5.3, we were able to invalidate the fourth hypothesis: although both systems use standard Arabic, more than half of the sentences were considered accurate. Furthermore, with both systems, for comprehensibility, our hypothesis remains invalidated for Lebanese and Egyptian vernacular-speaking patients; however, it is confirmed for Tunisian: less than half of the translations were considered fluent. With regards to dangerousness, our hypothesis stated that for both systems, at least 10% of the translations would have been dangerous. Our hypothesis is invalidated for BabelDr: only 6% of BabelDr's translations were considered dangerous, but 16% of Google Translate's questions were dangerous.

4.2.2.1 Adequacy

BabelDr

Table 4.11 shows the adequacy scores for BabelDr. According to a Tunisian-speaking person, out of 65 translations, 62% sentences are correct, 31% are ambiguous, 7% are mistranslations, and 0% nonsense. On the other hand, to a Lebanese-speaking person, for the same number of translations, 85% are correct, 9% are ambiguous, 6% are mistranslations, and 0% nonsense. Finally, to an Egyptian-speaking person, 78% translations are correct, 10% are ambiguous, 9% are mistranslations, and 3% are nonsense. The observed proportionate agreement for BabelDr was 0.53.

	Tunisian	**Lebanese**	**Egyptian**	**Total**
Correct	40 (62%)	55 (85%)	51 (78%)	**146**
Ambiguous	20 (31%)	6 (9%)	7 (10%)	**33**
Mistranslation	5 (7%)	4 (6%)	5 (9%)	**14**
Nonsense	0 (0%)	0 (0%)	2 (3%)	**2**

Table 4.11: BabelDr's adequacy: number of sentences.

Google Translate

Table 4.12 presents the adequacy scores for Google Translate. For a total of 96 translations, the Tunisian evaluator considered that 53% translations were correct, 34% ambiguous, 11% were mistranslations, and 2% were nonsense; the Lebanese evaluator found 79% correct translations, 9% ambiguous, 10% mistranslations, and 2% nonsense; the Egyptian evaluator found 67% correct translations, 12% ambiguous, 14% mistranslations, and 7% nonsense. The observed proportionate agreement for Google Translate was 0.44.

4.2.2.2 Comprehensibility

BabelDr

Table 4.13 illustrates the results for BabelDr regarding comprehensibility. For

	Tunisian	Lebanese	Egyptian	Total
Correct	51 (53%)	76 (79%)	64 (67%)	**172**
Ambiguous	32 (34%)	8 (9%)	11 (12%)	**51**
Mistranslation	11 (11%)	10 (19%)	14 (14%)	**35**
Nonsense	2 (2%)	2 (2%)	7 (7%)	**11**

Table 4.12: Google Translate's adequacy: number of sentences.

Tunisian, 21% of the sentences were considered fluent, 68% non-idiomatic, 8% contained a syntactic error, and 3% were incomprehensible; 100% of the sentences were evaluated as fluent in Lebanese; in Egyptian, 83% were considered fluent, 9% non-idiomatic, and 8% were considered incomprehensible. The observed proportionate agreement for BabelDr was 0.2.

	Tunisian	Lebanese	Egyptian	Total
Fluent	14 (21%)	65 (100%)	54 (83%)	**133**
Non-idiomatic	44 (68%)	0 (0%)	6 (9%)	**50**
Syntax error	5 (8%)	0 (0%)	0 (0%)	**0**
Incomprehensible	2 (3%)	0 (0%)	5 (8%)	**7**

Table 4.13: BabelDr's comprehensibility: number of sentences.

Google Translate

Table 4.14 illustrates the results for Google Translate regarding comprehensibility. For Tunisian, 38% of the translation is fluent, 42% are non-idiomatic, 13% contain a syntactic error, and 7% are incomprehensible; for Lebanese, 76% of the translations are fluent, 11% non-idiomatic, 5% contain a syntactic error, and 8% are incomprehensible; for Egyptian, 74% of the translations resulted fluent, 15% non-idiomatic, and 11% are incomprehensible. The observed proportionate agreement for Google Translate was 0.26.

	Tunisian	Lebanese	Egyptian	Total
Fluent	36 (38%)	73 (76%)	71 (74%)	**180**
Non-idiomatic	40 (42%)	10 (11%)	14 (15%)	**64**
Syntax error	13 (13%)	5 (5%)	0 (0%)	**18**
Incomprehensible	7 (7%)	8 (8%)	11 (11%)	**26**

Table 4.14: Google Translate's comprehensibility: number of sentences.

4.2.2.3 Dangerousness

Translators also evaluated the dangerous translations, in other words, the translations that can potentially mislead the health care professionals into making a wrong diagnosis. Table 4.15 illustrates the number of dangerous sentences with both systems. Only 6% of the translations with BabelDr were rated as dangerous, whereas 16% of the translations were considered dangerous with Google Translate.

	Tunisian	**Lebanese**	**Egyptian**	**Total**
BabelDr	1	2	1	**4 (6%)**
GT	3	7	6	**16 (16%)**

Table 4.15: Translators' dangerous sentences.

4.2.2.4 Translator's Kappa

Although translation students were asked to evaluate translations from the point of view of patients with basic knowledge of Arabic and who speak different Arabic vernaculars, it was still interesting to see to which degree they agree with each other and how well the system performs when employed with Arabic dialects. Hence, their inter-annotator agreement was calculated. BabelDr's inter-annotator agreement is very low for adequacy and comprehensibility (Light's Kappa for adequacy: 0.14; for comprehensibility: 0.014); while for Google Translate, the inter-annotator agreement is low for adequacy (Light's Kappa: 0.25) and very low for comprehensibility (Light's Kappa: 0.16) according to Landis & Loch [Landis and Koch, 1977].

4.2.3 Experts' evaluations

As mentioned, two experts were asked to evaluate both systems' accuracy and the dangerousness of the translations to assess the fifth hypothesis: from a pharmacist (domain expert) point of view, all of BabelDr's translations are accurate and none are dangerous, but not all of Google Translate's translations are accurate and at least 3% are dangerous. In this case, our hypothesis is invalidated for BabelDr: the two experts evaluated only 83% and 85% of the sentences as accurate. However, the hypothesis is confirmed with Google Translate: for both pharmacists, 79% of the translations are accurate. As far as dangerousness is concerned, our hypothesis remains invalidated: with both systems, none of the translations was considered dangerous.

4.2.3.1 Accuracy

BabelDr

Table 4.16 illustrates experts' evaluations of BabelDr. The first pharmacist rated 83% of the translations as correct, 14% as ambiguous, and 3% as mistranslations. The second pharmacist rated that 85% of the translations are correct, 9% are ambiguous, and 6% are mistranslations. The observed proportionate agreement for

BabelDr was 0.86.

Some of the translations that were rejected were "avez-vous une maladie chronique?" (do you have a chronic disease?) as in Arabic this was translated as "avez-vous une maladie aigue?" (do you have an acute illness?); another example of rejected translation was "avez-vous des problèmes gynécologiques?" (do you have any gynecological problems?) as in Arabic this was translated as "avez-vous des problèmes sexuels ou de procréation?" (do you have sexual or reproductive problems?), and "je vais m'occuper de vous aujourd'hui" (I will take care of you today) translated in Arabic as "je vais examiner votre situation aujourd'hui" (I will examine your situation today).

	Expert 1	**Expert 2**	**Total**
Correct	54 (83%)	55 (85%)	**109**
Ambiguous	9 (14%)	6 (9%)	**15**
Mistranslation	2 (3%)	4 (6%)	**6**
Nonsense	0 (0%)	0 (0%)	**0**

Table 4.16: BabelDr's adequacy: number of sentences.

Google Translate
Table 4.17 presents experts' evaluations of Google Translate. The first pharmacist rated 79% of the translations as correct, 12% as ambiguous, 4% as mistranslations, and 5% as nonsense. The second pharmacist rated that 79% of the translations are correct, 11% are ambiguous, 7% mistranslations, and 3% as nonsense. The observed proportionate agreement for Google Translate was 0.75.

Some of the rejected translations were "vous pouvez me dire où vous avez mal" (can you tell me where it hurts) which has been translated in Arabic as "You can tell me where it hurts 1 kilo", as we can see, "1 kilo" has been added at the end of the sentence which has been caused by the fact that the pharmacist left the microphone on and it was recording the Arabic audio playing for the patient. Other rejected translations were "est-ce que vous arrivez à avaler?" (can you swallow) which has been translated as "do you want to swallow?", and "vous avez mal quand vous faites pipi?" (does it hurt when you pee?) which has been considered as a mistranslation as it has been translated as "does your pee hurt?".

4.2.3.2 Dangerousness

According to the experts, none of the two systems produced dangerous translations which could lead to misleading answers from the patient.

4.2.3.3 Expert's Kappa

Inter-annotator agreement for BabelDr is moderate (Cohen's Kappa: 0.51), and it is low for Google Translate (Cohen's Kappa: 0.299) according to Landis & Loch

	Expert 1	Expert 2	Total
Correct	76 (79%)	76 (79%)	**152**
Ambiguous	11 (12%)	10 (11%)	**21**
Mistranslation	4 (4%)	7 (7%)	**11**
Nonsense	5 (5%)	3 (3%)	**8**

Table 4.17: Google Translate's adequacy: number of sentences.

[Landis and Koch, 1977].

4.2.3.4 Kappa: Tunisia

As an expert and a translator came from the same country (Tunisia) their Kappa score was calculated. Analyzing this aspect was quite interesting, as it shows how a language can become a barrier even between two people coming from the same country. Given that the pharmacist was only given adequacy, only this aspect could be measured. With both systems, the inter-annotator agreement was very low: for BabelDr, Cohen's Kappa was 0.103, and for GT, it was 0.18, according to Landis & Loch [Landis and Koch, 1977].

4.3 Conclusion

In this section, we presented the results of the data we collected. We started by presenting the results of the quantitative data (section 4.1). First, we showed the answers to the demographic questionnaire (4.1.1); secondly, we illustrated the diagnosis and triage questions of all three scenarios (4.1.2); thirdly, we presented the system usage (4.1.3) with a focus on the successful and unsuccessful interactions with the systems (4.1.3.1) and the time to complete the scenarios (4.1.3.2). We then proceeded with presenting the results of the qualitative data collected (4.2), which comprised three sections: results of the user satisfaction (4.2.1), the translator's evaluations (4.2.2), and the experts' evaluations (4.2.3).

Chapter 5

Discussion

The information gathered during our experiment and described in the previous chapter allows us to verify our five hypotheses (stated in section 1.2).

5.1 Triage questions and diagnoses

Hypothesis 1: Pharmacists can always reach a correct diagnosis with BabelDr as well as with Google Translate.

Our experiment results show that this hypothesis is true: pharmacists reached a correct diagnosis with both systems with all the scenarios.

Indeed, the scenarios were not complex and are pretty recurrent in pharmacy. However, it is worth mentioning that the pharmacist did not receive any information about where the pain was located and had to give a diagnosis based only on information collected during the tests. If we observe the tables presenting the most frequent questions in section 3.6.1.2, we can see that the questions taking into account the conception of time and if the patient has already taken medication are the most common. This indicates the two main flags that pharmacist use to make a diagnosis during a triage. Furthermore, during the tests, pharmacists revealed that asking closed questions was the exact opposite of what they are usually instructed to do in their daily practice. Typically, they should ask open questions to gather information from the patient's history. Unfortunately, BabelDr's design compels users to ask many questions to have only a yes or no answer. Hence, pharmacists often consulted the history of BabelDr to check if all the red flags had been covered and what questions were still missing. Furthermore, we discovered that the cystitis scenario caused some difficulties with the pain location, as Muslim women playing the role of patients had difficulties indicating their genitalia. BabelDr, which has pictographs, lacked one for this particular body part which would have helped the pharmacists locate the pain more quickly.

5.2 System usage

Hypothesis 2: BabelDr is the most usable system to perform triage in pharmacies compared to Google Translate with regards to successful interactions but not with regards to time.

The reason behind this hypothesis was that BabelDr, specifically developed for the medical domain, would have proposed more precise canonical questions that would have resulted in more successful interactions. However, by needing more manipulations than Google Translate (the pharmacist needs to select the right canonical question, select the right pictograph, and then come back to the home screen as seen in section 3.2), and given the fact that every pharmacist declared to have used Google Translate before (as seen in section 4.1.1), we expected that Google Translate would have allowed performing a faster triage even with a higher number of recognition errors.

Our hypothesis turned out to be false on both criteria stated. Google Translate proved to be the most efficient system for successful interactions. However, we have seen that since pharmacists do not understand the target language, it is not easy to define a successful interaction. By our definition of successful interaction in the case of Google Translate, pharmacists needed to play the translation in the system for the patient to hear. This was the case for 93% of the oralized questions with Google Translate. Successful interactions with BabelDr were considered those sent in their canonical form to translation. This was the case for 87% translations with BabelDr.

However, the fact that in BabelDr the pharmacists needed to validate the canonical question which might have been different from their oralized ones, may have played a role in rejecting the proposed canonical and hence in the percentage of successful interactions as well. Moreover, BabelDr's coverage is composed of a fixed set of sentences, which means that data is inserted manually and therefore is of higher quality. In contrast, Google Translate's system is data-driven, which means that the corpus on which it operates to generate the translations is not supervised, but it has more data to cover all the pharmacists' questions. Nevertheless, in the medical domain, having more precise and correct data is desirable, although there might be some gaps in the system coverage: a pharmacist can always try and formulate the question in another way. Whereas having much data which is not customized to the medical domain might yield inaccurate translations and wrong voice recognition. In fact, the reasons for rejecting the translations in Google Translate were speech recognition errors, and especially homophones seem to have caused the most difficulties for this system. Regarding BabelDr, the reasons for rejecting the translations were mainly due to out-of-coverage questions (either because they were out of domain or out of grammar).

With regards to time, we would have assumed that given the popularity of Google Translate and also the fact that all the participants declared having used it before, the triage would have gone faster with this system. However, both BabelDr and Google Translate proved fast at the same measure, invalidating our second hypothe-

sis. One reason why this happened might have been that BabelDr's system contains precise canonical questions, which made the oral questions more straightforward, leading to precise answers. Pharmacists revealed that in their daily practice they should ask open-ended questions to collect more data. In our scenarios, patients could mainly answer "yes" and "no". Therefore, the more the question was open-ended, the more time-consuming the triage became. In this aspect, we can affirm that BabelDr is the more efficient system in terms of time, especially if we consider that none of the pharmacists used the system before and could complete the scenarios at the same speed as with a system that had already been employed.

5.3 User Satisfaction

Hypothesis 3: Pharmacists prefer employing BabelDr rather than Google Translate to perform triage as BabelDr has been specifically developed for the medical domain.

The results of the qualitative data collected in section 4.2.1 have proven the hypothesis true. To test this hypothesis, we have collected data regarding ease of use, performance, and personal opinion.

With regards to ease of use, out of the eight criteria, BabelDr proved to be the best system in four cases. It was considered to be the system that presented better recognition results, the preferred one to interact with the patient, the one that allowed to ask enough questions to be sure of the decision, and that helped keep track of the questions. However, Google Translate was thought to be the easiest system to use, the easiest system to learn to use, the one that helped the most, and that allowed one to ask questions without feeling too constrained. The reasons for the preference for Google Translate could be the following: BabelDr forces the pharmacist do adapt their question to the canonical one BabelDr offers, whereas Google Translate allows the user to translate the exact speech recognition result. Also, given that participants declared to have already used Google Translate before, we can speculate that they felt that the system was easier to use thanks to their previous familiarity with the system.

As far as performance is concerned, BabelDr was the system that recognized the voice more easily, offered a trusted translation, and helped reach a conclusion. The reason behind this might have been the fact that participants were told during the introduction that BabelDr's translations were the results of the work of professional translators and interpreters, which is why they might have trusted the translations more and that their conclusions could be reached better. Nevertheless, Google Translated was the system that reacted the fastest to the requests. The reason behind this is that Google Translate's microphone is better than BabelDr's microphone.

With respect to personal opinion, BabelDr was thought to be the system that can improve the triage process for allophone patients, that can easily be integrated into the pharmaceutical practice, and that can help pharmacists save time. All these preferences might correspond to the doctor's perceptions stated in section 1:

pharmacists might be skeptical as well about employing a system that has not been trained for the medical domain.

5.4 Translators' Evaluations

Hypothesis 4: Since both systems employ standard Arabic, only a couple of the questions will be accurate, comprehensibility scores will be low, and at least 10% of the translations will be dangerous with BabelDr and Google Translate for patients speaking vernacular Arabic and with little understanding of Standard Arabic.

This hypothesis comprised three aspects to be tested: accuracy, comprehensibility, and dangerousness with both systems in three different vernacular Arabic languages. Our hypothesis is invalidated in all aspects apart from the fluency of Tunisian. Only in this dialect less than half of the translations were considered fluent.

Starting with accuracy, more than half of the translations were correct. With BabelDr, Lebanese had the highest scores (85%), followed by Egyptian (78%) and Tunisian (62%). With Google Translate, Lebanese presented the highest scores (79%), followed by Egyptian (67%) and Tunisian (53%). We can observe that BabelDr's Translations are more accurate than Google Translate's for all three vernaculars.

With regards to comprehensibility, Lebanese and Egyptian had high scores, whereas Tunisian presented the lowest. This can be due to the difference between this dialect and Standard Arabic. For Tunisian, in fact, our hypothesis is true: comprehensibility scores are low with both systems. More in detail, with BabelDr, 21% of the Tunisian translations were fluent, 100% of the Lebanese ones were fluent, while 83% of the Egyptian translations were considered fluent. With Google Translate, comprehensibility scores were higher for Tunisian: 38% of the translations were considered fluent. But scores were lower for the other two vernaculars: 76% and 74% of the translations with Lebanese and Egyptian were considered fluent. Our fourth hypothesis stays invalidated overall.

Lastly, as far as dangerousness is concerned, we hypothesized that at least 10% of the translations were dangerous with both systems. This hypothesis is invalidated for BabelDr but valid for Google Translate. In fact, 6% of BabelDr's translations were considered dangerous, whereas 16% of Google Translate's translations were considered dangerous. More in detail, Lebanese proved to be the vernacular for which translations with both systems were the most dangerous: the translator evaluated 3% of the translations as dangerous with BabelDr, and 7% as dangerous with Google Translate. As we can see, BabelDr is the best system in this regard, producing 10% less dangerous translations than Google Translate even when it comes to dialects. We think this might be due to the fact that the professional translators who took care of the coverage for this language produced culturally more sensitive translations that can be understood by a wider group of patients.

5.5 Experts' Evaluations

Hypothesis 5: From a pharmacist (domain expert) point of view, all of BabelDr's translations are accurate, and none are dangerous, but not all of Google Translate's translations are accurate, and at least 3% are dangerous.

In this case, the hypothesis needs to be divided into two parts: the accuracy and dangerousness of both systems.

As far as accuracy is concerned, our hypothesis is invalidated with BabelDr. In fact, only some translations were considered accurate (only 83% and 85%). Nevertheless, the hypothesis remains true for Google Translate: 79% and 76% of the translations are considered accurate. We can still affirm that although not all of BabelDr's sentences are accurate, they are still more accurate than Google Translate's; hence, the former system is a better fit for the pharmacy domain when it comes to accuracy of terms and formulations. This might be due to the fact that BabelDr's system contains only the translations of experts and has been specifically developed for the medical domain.

With regards to dangerousness, our hypothesis is invalidated for Google Translate: in fact, none of the translations are considered dangerous with this system. However, the hypothesis remains true for BabelDr: none of the translations are dangerous with this system. It comes with comfort the fact that although Google Translate is being used in medical settings (while not having been developed for the medical context), it does not yield dangerous translations with regards to the three scenarios we tested. We think this might be due to the fact that the system has a good amount of data for this language pair that allows for high-quality results.

Chapter 6

Conclusion

This last chapter will first present a synthesis of the results of our research; secondly, it will illustrate the limits of our study with research leads for future work.

6.1 Summary of the study

Our study aimed to determine whether BabelDr was better than Google Translate at performing triage in a pharmacy setting. We have formulated five hypotheses which guided our research. According to our hypotheses, pharmacists can always reach a correct diagnosis with BabelDr as well as with Google Translate; BabelDr is the most usable system to perform triage in pharmacists compared to Google Translate in terms of successful interactions, but not with regards to time; pharmacists prefer employing BabelDr rather than Google Translate to perform triage as BabelDr has been specifically developed for the medical domain; translations with both systems would not have been accurate nor comprehensible, and at least 10% of the translations would have been dangerous for patients speaking Arabic vernaculars such as Egyptian, Lebanese, and Tunisian. Our last hypothesis stated that from an expert point of view, all BabelDr's translations are accurate and none dangerous, but not all Google Translate's translations are accurate, and at least 3% are dangerous.

We carried out eight tests and two evaluation sessions which have allowed us to test these hypotheses. We first chose two systems: BabelDr, a flexible phraselator that employs speech recognition, and Google Translate, a neural machine translation system. In the first part of our method of experimentation, we asked eight pharmacists to perform triage with both systems. They were then asked to fill in a user satisfaction questionnaire which allowed us to collect their opinion. In the second part, we asked master-level translation students and pharmacists to evaluate the translations to collect data about the quality of these.

During the tests, pharmacists reached a correct diagnosis with both systems for every scenario confirming our first hypothesis. The scenarios given were quite simple and recurrent in a pharmacy. However, unlike other studies conducted following this method, pharmacists were given no information as to the location of the pain nor the symptoms. We have seen that this represented quite an issue in the cystitis scenario, where Muslim women hesitated in indicating the location of the pain. We can see how both systems can only help a little in this regard. On the one side, BabelDr

does not offer a pictograph for this body part; on the other, Google Translate does not offer any help. It would be good for BabelDr to add one pictograph relative to female genitalia to overcome this barrier. As for Google Translate, which has yet to be developed specifically for the medical domain, we might suggest adding pictographs as well or images with body parts.

Regarding the successful interactions, we have seen that Google Translate turned out to be the best system, with 93% of successful interactions compared to BabelDr's 87%, disproving our second hypothesis. The reasoning behind our hypothesis was that BabelDr has been developed for the medical domain. Therefore recognition results would have been more precise and domain-specific. However, the fact that this system does not offer the precise voice recognition result when sending the question to translation but rather it offers a canonical sentence has led to pharmacists rejecting more recognition results than with Google Translate.

As far as time is concerned, our hypothesis remains invalidated. Although pharmacists declared in section 4.1.1 that they all already used Google Translate, the time taken to complete the scenarios were almost the same. Moreover, if we take into account that BabelDr takes more manipulation than Google Translate (activating the microphone, selecting the canonical question, selecting a pictograph, and then coming back to the main page) and that this was the first time that pharmacists employed this system, we can suppose that BabelDr has allowed the pharmacists to reach a correct diagnosis faster than Google Translate. In fact, for this system, the manipulations are less: the user only needs to activate the microphone and then press the button to reproduce the translation.

We analyzed the results of the user satisfaction questionnaire. The questionnaire covered three main aspects: ease of use, performance, and personal opinion. BabelDr proved to be the preferred system on most of the criteria, confirming our third hypothesis. Pharmacists considered BabelDr to be the system that presented better recognition results. This may be since questions when sent to translation, are always accompanied by pictographs. BabelDr's feature history allows pharmacists to keep track of the questions, and this might be the reason for them preferring BabelDr to Google Translate in this aspect. Pharmacists also trusted BabelDr's translation more than Google Translate's. We reckon this is due to the fact that during the introduction to the tests, we made it clear that BabelDr's translations were the result of the work of professional translators and interpreters. However, the results were more nuanced than this. Also, some of Google Translate's features were considered more appealing: it was considered the easiest system to use and learn, and it allowed to ask questions without being too constrained. Clearly, it is important to look at the bigger picture and to take into account the nuanced opinions of pharmacists in order to improve BabelDr's current features.

We carried out a series of test evaluations with master-level translation students and pharmacists. We have seen the importance of the choice of the right judges in section 3.5.3.

In our case, for the linguistic evaluation, we deemed it important to have the

point of view of patients who have little understanding of standard Arabic, as we have seen in section 3.3 that illiteracy percentages are still high in Arabic-speaking countries. Our results showed that most of the translations were accurate for three vernacular Arabic (Tunisian, Egyptian, and Lebanese) with both systems. They were comprehensible for Egyptian and Lebanese but not for Tunisian. BabelDr also proved to be the system that presented with the smaller amount of dangerous translations (only 6%), whereas Google Translate's translations proved to be almost three times as dangerous, with 16% of them rated as such. Hence, we can affirm that BabelDr is the better system for all three vernaculars in all three aspects (accuracy, comprehensibility, and dangerousness) apart from comprehensibility for Tunisian.

Experts evaluated the translations for standard Arabic, as we wanted to make sure that the terminology was correct for the pharmacy domain. None of the systems provided 100% accurate translations. Invalidating our fifth hypothesis as far as accuracy is concerned. However, BabelDr's translations were considered more accurate (83% and 85% for BabelDr against 79% and 76% for Google Translated). The fifth hypothesis remains invalidated for dangerousness as well: both systems provided translations that were not dangerous at all. Hence, despite the fact that Google Translate has not been developed for the medical domain, thanks to its immense data, it is a valid alternative for translation in the pharmacy sector but BabelDr should be the preferred system that pharmacists should use.

6.2 Limitations of the study and perspectives

Our study has its own number of limits that can be avoided in future tests.

First of all, the limited number of pharmacists does not allow us to generalize and to reach statistically significant results. In fact, it would be interesting to see how a higher number of pharmacists might feel about the two systems. Moreover, it would also be preferable not to have pharmacists from only one pharmacy but from pharmacies all across Geneva and also to have more data on the languages and the respective vernaculars. Not only we had a small number of participants, but the number of scenarios tested was small. For BabelDr we also made sure beforehand that all questions were present. Future studies might want to investigate a broader range of scenarios in order to see which domains need more coverage in pharmacies for BabelDr.

A limitation of our study was that participants playing the part of patients understood French. Therefore, they were able to hear the source question before it went to translation. Hence, they compared the source to the target before giving their answers. For future research, it might be useful to have a participant playing the role of the patient with no understanding of French so that scenarios and tests yield unbiased results. Furthermore, testing the application in a real pharmacy scenario with real patients might also give an idea of the advantages and limits of both systems.

Another limitation of the study might have been the subjectivity of the evaluators. On the one hand, choosing humans was necessary to be sure that translations

were accurate and fluent from a human perspective, which could not have been possible if we relied on automated metrics which would have compared the output of the systems to a human golden standard and penalised the translations that, although correct, were too different from the latter. In our case, for the three types of vernacular Arabic, given that these are only spoken languages, automated metrics would not have been possible. We needed evaluators who could also understand the source language in order to measure accuracy. However, by using the same linguistic evaluators for comprehensibility, our results might have been invalidated, as evaluators might be influenced by the accuracy evaluations when assessing comprehensibility [Koehn and Senellart, 2010].

Our study has brought up the issue of standard Arabic not being comprehensible by everyone in the Arabic-speaking world, and thanks to our research, the new version of BabelDr has added three different vernacular Arabic to their language coverage (Algerian, Moroccan, and Tunisian Arabic). This means that nowadays patients can communicate in their dialect when visiting the doctor or the pharmacist.

This study is a small first step in researching applications that can be used in a pharmacy setting to overcome language barriers; it does not claim to present any generalized results and further studies are needed. However, we hope this research will inspire more studies in the domain to have a tool that fits most pharmacists' needs and patients' requests. Furthermore, in an increasingly globalized world, where we face humanitarian crises that drive more and more people to flee their countries, improving a tool like BabelDr or developing something similar might bring us closer to more equitable access to health care while still helping to reduce costs and unburdening emergency departments.

Bibliography

Ann D Bagchi, Stacy Dale, Natalya Verbitsky-Savitz, Sky Andrecheck, Kathleen Zavotsky, and Robert Eisenstein. Examining effectiveness of medical interpreters in emergency departments for spanish-speaking patients with limited english proficiency: results of a randomized controlled trial. *Annals of emergency medicine*, 57(3):248–256, 2011.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

Aaron Bangor, Philip Kortum, and James Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3): 114–123, 2009.

Gillian Bartlett, Régis Blais, Robyn Tamblyn, Richard J Clermont, and Brenda MacGibbon. Impact of patient communication problems on the risk of preventable adverse events in acute care settings. *Cmaj*, 178(12):1555–1562, 2008.

Luisa Bentivogli, Mauro Cettolo, Marcello Federico, and Christian Federmann. Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment. In *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT 2018)*, pages 62–69, 2018.

Alexander Bischoff and Patricia Hudelson. Access to healthcare interpreter services: where are we and where do we need to go? *International journal of environmental research and public health*, 7(7):2838–2844, 2010.

Alexander Bischoff, Claude Tonnerre, Ariel Eytan, Martine Bernstein, and Louis Loutan. Addressing language barriers to health care, a survey of medical services in switzerland. *Sozial-und Präventivmedizin*, 44(6):248–256, 1999.

Pierrette Bouillon and Hervé Spechbach. Babeldr: a web platform for rapid construction of phrasebook-style medical speech translation applications. In *19th annual conference of the European Association for Machine Translation (EAMT)*, 2016.

Pierrette Bouillon, Johanna Gerlach, Hervé Spechbach, Nikolaos Tsourakis, and Ismahene Sonia Halimi Mallem. Babeldr vs google translate: A user study at geneva university hospitals (hug). In *20th Annual Conference of the European Association for Machine Translation (EAMT)*, 2017.

Pierrette Bouillon, Johanna Gerlach, Jonathan Mutal, Nikos Tsourakis, and Hervé Spechbach. A speech-enabled fixed-phrase translator for healthcare accessibility. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 135–142, 2021.

Valérie Boujon, Pierrette Bouillon, Hervé Spechbach, Johanna Gerlach, and Irene Strasly. Can speech-enabled phraselators improve healthcare accessibility? a case study comparing babeldr with medibabble for anamnesis in emergency settings. In *Proceedings of the 1st Swiss Conference on Barrier-free Communication*, 2018.

Michael Bradshaw, Sandra Tomany-Korman, and Glenn Flores. Language barriers to prescriptions for patients with limited english proficiency: a survey of pharmacies. *Pediatrics*, 120(2):e225–e235, 2007.

John Brooke. Sus: a "quick and dirty'usability. *Usability evaluation in industry*, 189(3), 1996.

Chris Callison-Burch, Cameron Shaw Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, 2007.

Dwayne TS Chang, IA Thyer, Dickon Hayne, and DJ Katz. Using mobile technology to overcome language barriers in medicine. *The Annals of The Royal College of Surgeons of England*, 96(6):e23–e25, 2014.

Eirini Chatzikoumi. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161, 2020.

Adam L Cohen, Frederick Rivara, Edgar K Marcuse, Heather McPhillips, and Robert Davis. Are language barriers associated with serious medical events in hospitalized pediatric patients? *Pediatrics*, 116(3):575–579, 2005.

Confédération. Statistique en matière d'asile, novembre 2022. "`https://www.sem.admin.ch/sem/fr/home/publiservice/statistik/asylstatistik/archiv/2022/11.html`", accessed December 30, 2022.

Confédération. Remise simplifiée des médicaments soumis à ordonnance, october 2022. "`https://www.bag.admin.ch/bag/fr/home/medizin-und-forschung/heilmittel/abgabe-von-arzneimitteln.html`", accessed January 8, 2023.

Janine Dahinden, Joëlle Moret, Jean-Christophe Loubier, Véronique Meffre, and Dominik Weber. Vers une meilleure communication: Coopération avec les réseaux de migrants. Technical report, Université de Neuchâtel, 2009.

Shelagh Daly. Medibabble translator app. *Nursing Standard (2014+)*, 28(42):33, 2014.

Carol C Diamond, Farzad Mostashari, and Clay Shirky. Collecting and sharing data for population health: a new paradigm. *Health affairs*, 28(2):454–466, 2009.

Glenn Flores, M Barton Laws, Sandra J Mayo, Barry Zuckerman, Milagros Abreu, Leonardo Medina, and Eric J Hardt. Errors in medical interpretation and their potential clinical consequences in pediatric encounters. *Pediatrics*, 111(1):6–14, 2003.

Caroline Free, J Green, V Bhavnani, and A Newman. Bilingual young people's experiences of interpreting in primary care: a qualitative study. *British Journal of General Practice*, 53(492):530–535, 2003.

Johanna Gerlach, Pierrette Bouillon, Rovena Troqe, Sonia Halimi, and Hervé Spechbach. Patient acceptance of translation technology for medical dialogues in emergency situations. *Translating Crises*, page 253, 2022.

Emina Hadziabdic and Katarina Hjelm. Arabic-speaking migrants' experiences of the use of interpreters in healthcare: a qualitative explorative study. *International journal for equity in health*, 13(1):1–12, 2014.

Sonia Halimi and Pierrette Bouillon. Google translate and babeldr in community medical settings: Challenges of translating into arabic. In *Arabic Translation Across Discourses*, pages 27–44. Routledge, 2019.

Sonia Asmahène Halimi, Razieh Azari, Pierrette Bouillon, and Hervé Spechbach. A corpus-based analysis of medical communication: Euphemism as a communication strategy for context-specific responses. In *Corpus Exploration of Lexis and Discourse in Translation*, pages 1–25. Routledge, 2021.

HUG. Babeldr : Mieux se comprendre à l'hôpital, 2022.

Tomasz Kamusella. The arabic language: A latin of modernity? *Journal of Nationalism, Memory & Language Politics*, 11(2):117–145, 2017.

Dorothy Kenny. Human and machine translation. *Machine translation for everyone: Empowering users in the age of artificial intelligence*, 18:23, 2022.

Geoffrey S Koby, Paul Fields, Daryl R Hague, Arle Lommel, and Alan Melby. Defining translation quality. *Tradumàtica*, (12):0413–420, 2014.

Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

Philipp Koehn. *Neural Machine Translation*. Cambridge University Press, 2020. doi: 10.1017/9781108608480.

Philipp Koehn and Jean Senellart. Convergence of translation memory and statistical machine translation. In *Proceedings of AMTA Workshop on MT Research and the Translation Industry*, pages 21–31, 2010.

J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

Alon Lavie. Evaluating the output of machine translation systems. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Tutorials*, 2010.

Annim Mohammad, Bandana Saini, and Betty Bouad Chaar. Pharmacists' experiences serving culturally and linguistically diverse patients in the australian community pharmacy setting. *International Journal of Clinical Pharmacy*, pages 1–11, 2021.

Jonathan David Mutal, Johanna Gerlach, Pierrette Bouillon, and Hervé Spechbach. Ellipsis translation for a medical speech to speech translation system. In *22nd Annual Conference of the European Association for Machine Translation (EAMT)*, 2020.

Joseph Olive, Caitlin Christianson, and John McCary. *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation*. Springer Science & Business Media, 2011.

World Health Organization. Joint fip/who guidelines on good pharmacy practice: standards for quality of pharmacy services. `https://www.who.int/medicines/areas/quality_safety/quality_assurance/FIPWHOGuidelinesGoodPharmacyPracticeTRS961Annex8.pdf`, 2011. (Accessed on 08/11/2021).

Anita Panayiotou, Anastasia Gardner, Sue Williams, Emiliano Zucchi, Monita Mascitti-Meuter, Anita MY Goh, Emily You, Terence WH Chong, Dina Logiudice, Xiaoping Lin, et al. Language translation apps in health care settings: Expert opinion. *JMIR mHealth and uHealth*, 7(4):e11316, 2019.

Sumant Patil and Patrick Davies. Use of google translate in medical communication: evaluation of accuracy. *Bmj*, 349, 2014.

Thierry Poibeau. *Babel 2.0: où va la traduction automatique?* Odile Jacob, 2019.

Alvi A Rahman. Rising up to the challenge: strategies to improve health care delivery for resettled syrian refugees in canada, 2016.

Humbel Ruth. Place des pharmacies dans les soins de base. "`https://www.parlament.ch/en/ratsbetrieb/suche-curia-vista/geschaeft?AffairId=20123864`", 2012.

David LB Schwappach, Carla Meyer Massetti, and Katrin Gehring. Communication barriers in counselling foreign-language patients in public pharmacies: threats to patient safety? *International journal of clinical pharmacy*, 34(5):765–772, 2012.

Mark Seligman and Mike Dillinger. Automatic speech translation for healthcare: some internet and interface aspects. In *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence (TIA-13)*, pages 28–30, 2013.

Sharaf Sheik-Ali, Naeem Dowlut, and Greg McConaghie. Breaking down language barriers with technology. *The Bulletin of the Royal college of Surgeons of england*, 98(3):138–140, 2016.

David Silvera-Tawil, Courtney Pocock, DanaKai Bradford, Andrea Donnell, Jill Freyne, Karen Harrap, Sally Brinkmann, et al. Enabling nurse-patient communication with a mobile app: Controlled pretest-posttest study with nurses and non–english-speaking patients. *JMIR Nursing*, 4(3):e19709, 2021.

Hervé Spechbach, Johanna Gerlach, Sanae Mazouri Karker, Nikos Tsourakis, Christophe Combescure, Pierrette Bouillon, et al. A speech-enabled fixed-phrase translator for emergency settings: Crossover study. *JMIR medical informatics*, 7 (2):e13167, 2019.

Dominik Stampfli, Birgit Alexandra Winkler, Simona Berardi Vilei, and Andrea M Burden. Assessment of minor health disorders with decision tree-based triage in community pharmacies. *Research in Social and Administrative Pharmacy*, 2021.

Rachel M Taylor, Nicola Crichton, Beki Moult, and Faith Gibson. A prospective observational study of machine translation software to overcome the challenge of including ethnic diversity in healthcare research. *Nursing Open*, 2(1):14–23, 2015.

TS Tullis and JN Stetson. A comparison of questionnaires for assessing website usability. usability professionals association conference. *Minnesota, USA2004*, 2004.

UniversalPharmacist, 2022. URL `http://www.u-pharmacist.es/aplicacion-web.htmlm`.

Charles A Vincent and Angela Coulter. Patient safety: what about the patient? *BMJ Quality & Safety*, 11(1):76–80, 2002.

Ryckie G Wade. Try google translate to overcome language barriers. *Bmj*, 343, 2011.

Appendix A

Appendix

A.1 Scenarios

Les trois scénarions prévus

Scénario 1 – Cystite chez une femme de 25 ans

- Miction (l'action d'uriner) difficile, douloureuse (+) (symptômes typiques que la patiente reconnaît)
- Mictions (l'action d'uriner) fréquentes, en petite quantité (+)
- Début des symptômes : depuis 24 heures
- Prurit vaginal (-)
- Pertes vaginales (-)
- Premier épisode (première fois) (-)
- Dernière infection urinaire : > 6 mois
- Fréquence des infections urinaires : max. 1 fois par année
- Sang dans les urines (-)
- Douleurs bas ventre (-)
- Douleurs aux flancs (-)
- Enceinte (-)
- Maladie chronique (-), bonne santé (+)
- Vomissements (-)
- Fièvre, frissons (-)
- Médecin de famille (-)
- Relations sexuelles à risque (-)

Résolution : remise d'un médicament = un antibiotique

NB. Le patient va devoir payer cette prise en charge, car non prise en charge par de nombreux assureurs maladie.

Scenario 2 – Mal de tête chez un homme/femme de 33 ans qui souhaite un médicament pour la soulager et lui permettre de poursuivre son travail au bureau jusqu'à la fin de la journée, journée chargée

- Douleur unilatérale (d'un seul coté)(-)
- Rhinorrhée (écoulement de nez) (-)
- Fièvre (-)
- Perte de poids (-)
- Vomissements (-)
- Douleurs/pression dans la face (-)
- Atteinte de l'odorat (-)
- Céphalées (douleur à la tête) inhabituelles (-)
- Douleur déclenchée par un effort (-), un changement de postions (-), valsalva (consiste à expirer de l'air vers les trompes d'Eustache, en ayant la bouche fermée et le nez pincé (effort d'expiration forcée à glotte fermée pendant une quinzaine de secondes) (-).

- Douleur/pression qui vient et qui part (+) en casque (maux de tête touchant les deux côtés du crâne) (+)
- Apparition de la douleur rapide/aiguë/intensité explosive (-)
- Fréquence de céphalées : rare (1 fois /6 mois)
- Durée des douleurs < 72 heures (+)
- Si femme, grossesse (-)
- Bon état général (+)
- Comorbidités (ex. hypertension (maladie caractérisée par une pression artérielle trop élevée), troubles neurologiques, troubles de l'hémostase (arrêt d'une hémorragie) connus, glaucome (maladie dégénérative du nerf optique qui entraîne une perte progressive de la vision commençant tout d'abord en périphérie et progressant graduellement vers le centre) (-), aucun trouble du sommeil) (-)
- Traumatisme crânien dans les 3 derniers mois (-)
- Prise de médicaments ces derniers jours (-)
- Prise habituelle de médicaments (-)

Résolution : remise d'un médicament : paracétamol ; accompagnée de recommandations : sommeil suffisant et de qualité; activité physique régulière ; exercice de détente. Consulter le médecin si les maux de tête ne s'améliorent pas ou s'aggravent dans les 48 heures

Scenario 3 – Toux chez un homme/femme de 21 ans, période d'examen, la toux le.la réveille la nuit, aimerait mieux dormir et bien se reposer

- Durée des symptômes : 3 jours
- Expectorations (l'expulsion des sécrétions produites par les voies aériennes, cracher), toux productive, toux grasse (-)
- Rhume (+)
- Fièvre (-)
- Otite (infection et inflammation de l'oreille (-)
- Etat général (bon)
- Tabac (-)
- Comorbidités (ex. asthme, diabète) (-)
- Difficultés respiratoires (-)
- Difficultés à avaler (-)
- Douleurs thoraciques (-)
- Prise de médicaments habituels (vitamines ; pilule contraceptive pour la femme) (+)
- Douleurs musculaires (+)
- Perte de goût (-), odorat (-)
- Peine à respirer (-)
- Diarrhées (-)
- -lésions cutanées (-)
- Vaccinés pour le covid-19 deux doses (+)

- Contact avec personnes qui ont le covid-19 (-)
- Voyage dans les 14 jours précédents (-)

Résolution : possible infection virale, et toux induite par le rhume; solution saline et/ou vasoconstricteur local pour soigner le rhume (max. durant 5 jours), sirop à base de butamirate pour calmer la toux avant le coucher. Monitorer la température ; en cas de fièvre, si la toux devient productive, ou si les symptômes perdurent malgré le traitement, revoir son pharmacien ou consulter un médecin.

A.2 Instructions for the translation evaluations

25.01.2022

ÉVALUATIONS

1. INTRODUCTION

Pour les évaluations, vous avez reçu deux documents Excel. Chaque document a deux feuilles *Adequacy* et *Comprehensibility* et il présente des phrases en français avec leur traduction en arabe. L'ordre des phrases est aléatoire.

A. Tâches

i. Je vais vous demander d'évaluer chaque traduction selon des critères précis qui vont être décrits dans la section suivante ;

ii. Vous avez des origines différentes, donc je vous demanderais d'évaluer les traductions selon le « type » d'arabe que vous parlez : par exemple, j'aimerais savoir si un-e patient-e qui parle l'arabe maghrébin (ou bien l'arabe du golf ou l'arabe du moyen orient) pourrait comprendre la traduction proposée. Cette partie est très importante, donc je vous demande de ne pas évaluer l'arabe standard, mais plutôt de vous concentrer sur comment l'arabe de la traduction proposée serait aperçue par quelqu'un qui ne parle pas l'arabe standard, ou qui le connait très peu.

2. ADEQUACY

Dans cette feuille je vous demanderais de faire deux choses :

i. **D'évaluer si la traduction a la même signification que la phrase en français.** Le but n'est pas d'avoir une traduction précise, mot par mot, mais d'avoir le même sens et surtout une phrase qui respecte la culture du patient.

Donc, si par exemple la question en français demande :

Montrez-moi avec la main où se situe la douleur

et en arabe on traduit avec

Montrez-moi avec votre doigt où se situe la douleur

la traduction est tout à fait correcte car cela ne change rien pour le patient, qui doit indiquer où se situe la douleur.

La même chose dans le cas de

Avez-vous de la température ?

si on traduit avec

Avez-vous de la fièvre ou *Êtes-vous fiévreux ?*

La traduction est correcte car le concept est le même (on demande au patient s'il a la fièvre).

Cependant (je vous le présente ici avec un exemple anglais), si on traduit « nausea » avec « dizziness », c'est évidemment faux, car les deux ne sont pas la même chose.

Je vous prie de ne pas tenir compte si la traduction est littérale, mais plutôt si la traduction est pertinente, et de réfléchir du point de vue du patient.

ii. **Évaluer la « dangerousness ».** Le but de cette évaluation et d'identifier les traductions qui veulent dire totalement autre chose par rapport à la phrase source, celles qui dans un domaine médical sont donc dangereuses car elles pourraient aboutir à des réponses fausses de la part du patient et donc à donner de fausses indications au médecin.

Je vous donne ici quelques exemples :

Prenez-vous des médicaments?

هل يتعاطى **المخدرات؟** (Do you take narcotics?)

Vous avez mal dans quel endroit?

ما مكان في خطأ على كنت (You were wrong)

Afin de ne pas vous faire perdre du temps, je vous demande juste de noter que les phrases qui sont « dangereuses », pas besoin de noter les phrases correctes ou pas dangereuses.

3. COMPREHENSIBILITY

Dans cette feuille, vous allez trouver juste la liste des phrases en arabe que vous devrez évaluer. Si la phrase est correcte au niveau sémantique et syntaxique, alors la traduction est de bonne qualité. Essayez de ne pas noter les phrases subjectivement afin d'éviter d'avoir des évaluations qui sont très différentes pour des traductions correctes.

Je vous rappelle que votre participation sera rémunérée par la Doyenne à travers des bons cadeau (FNAC).

N'hésitez pas à me contacter en cas de doutes ou questions.

Je vous remercie de votre participation et votre travail.

Rebeka

A.3 Corpus BabelDr: Adequacy

Adequacy

	Score
correct	4
ambigous	3
mistranslation	2
nonsense	1

ID	Source	Translation	Score	Dangerous
1	avez-vous des démangeaisons ?	هَلْ تَشْكِينَ مِنَ الْحَكَّة ؟		
2	la douleur est-elle comme une brûlure ?	هَلْ الأَلَمُ يُشْبِهُ الْحُرْقَة ؟		
3	prenez-vous un autre médicament ?	هَلْ تَتَنَاوَلُ دَوَاءً آخَرَ ؟		
4	avez-vous mal ailleurs ?	هَلْ تَشْعُرِينَ بِأَلَمٍ فِي أَمَاكِنَ أُخْرَى مِنَ الْجِسْمِ ؟		
5	avez-vous des allergies connues ?	هَلْ تَمَّ تَشْخِيصُ أَيَّةِ حَسَاسِيَّةٍ لَدَيْك ؟		
6	avez-vous de la fièvre ?	هَلْ لَدَيْك الْحُمَّى ؟		
7	avez-vous des symptômes particuliers en ce moment ?	هَلْ تُلَاحِظُ ظُهُورَ أَعْرَاضٍ مُعَيَّنَةٍ الآنَ ؟		
8	pouvez-vous me montrer avec le doigt où est la douleur ?	هَلْ يُمْكِنُكِ الإِشَارَةُ بِالأَصْبَعِ إِلَى مِنْطَقَةِ الأَلَمِ ؟		
9	avez-vous une toux sèche ?	هَلْ تَشْكِينَ مِنْ سُعَالٍ جَافٍ ؟		
10	avez-vous une maladie chronique ?	هَلْ تُعَانِينَ مِنْ مَرَضٍ حَاذ ؟		
11	êtes-vous enceinte ?	هَلْ أَنْتِ حَامِل ؟		
12	la douleur est-elle intense ?	هَلْ الأَلَمُ شَدِيدٌ ؟		
13	depuis combien de jours ?	مُنْذُ كَمْ يَوْم ؟		
14	bonjour	مَرْحَبًا		
15	avez-vous mal quand vous urinez ?	هَلْ تَشْعُرِينَ بِأَلَمٍ عِنْدَ التَّبَوُّلِ ؟		
16	quelle est l'intensité de la douleur sur une échelle de zéro à dix, zéro étant le minimum et dix le maximum ?	مَاهِيَ شِدَّةُ الأَلَمِ عَلَى مِقْيَاسِ دَرَجَاتٍ مِنْ 0 إِلَى 10 عِلْمًا أَنَّ 0 يَعْنِي غِيَابَ الأَلَمِ وَ 10 الْحَدَّ الأَقْصَى لِلأَلَم ؟		
17	avez-vous pris un médicament ?	هَلْ تَنَاوَلْتِ دَوَاءً مَا ؟		
18	avez-vous reçu un coup à la tête ?	هَلْ تَعَرَّضْتَ لِضَرْبَةٍ فِي الرَّأْسِ ؟		
19	depuis combien de jours avez-vous mal ?	مُنْذُ كَمْ يَوْم تَشْعُرُ بِالأَلَمِ ؟		
20	quel âge avez-vous ?	كَمْ عُمْرُك ؟		
21	avez-vous mal aux épaules ?	هَلْ تَشْعُرُ بِأَلَمٍ فِي الْكَتِفَيْنِ ؟		
22	avez-vous des difficultés à avaler du liquide ?	هَلْ تَجِدُ صُعُوبَةً فِي شُرْبِ السَّوَائِل ؟		
23	avez-vous d'autres problèmes de santé ?	هَلْ لَدَيْك مَشَاكِلُ صِحِّيَّةٌ أُخْرَى ؟		
24	prenez-vous un médicament ?	هَلْ تَتَنَاوَلُ دَوَاءً مَا ؟		
25	avez-vous des difficultés à avaler ?	هَلْ تَجِدِينَ صُعُوبَةً فِي الْبَلْعِ ؟		
26	depuis combien de jours toussez-vous ?	مُنْذُ كَمْ يَوْم وَأَنْتِ تَسْعُلِينَ ؟		
27	avez-vous d'autres plaintes ?	هَلْ تَشْكِينَ مِنْ شَيْءٍ آخَرَ ؟		
28	depuis combien de jours avez-vous mal ?	مُنْذُ كَمْ يَوْم تَشْعُرِينَ بِالأَلَمِ ؟		
29	je vais m'occuper de vous aujourd'hui	سَأَتَوَلَّى النَّظَرَ فِي حَالَتِكِ الْيَوْمَ		
30	avez-vous pris des médicaments contre la douleur aujourd'hui ?	هَلْ تَنَاوَلْتِ أَدْوِيَةً مُسَكِّنَةً لِلأَلَمِ الْيَوْمَ ؟		
31	avez-vous souvent des migraines ?	هَلْ غَالِبًا مَا تُصَابُ بِالصُّدَاعِ النِّصْفِي ؟		
32	avez-vous mal quelque part en ce moment ?	هَلْ تَشْعُرِينَ بِأَلَمٍ فِي مَكَانٍ مَا الآن ؟		
33	toussez-vous?	هَلْ تَسْعُلِينَ ؟		
34	avez-vous pris votre tension ?	هَلْ قِسْتَ ضَغْطَ دَمِكَ ؟		
35	êtes-vous allergique à des médicaments ?	هَلْ لَدَيْك حَسَاسِيَّةٌ مِنْ بَعْضِ الأَدْوِيَة ؟		
36	avez-vous pris un médicament aujourd'hui ?	هَلْ تَنَاوَلْتَ دَوَاءً مَا الْيَوْمَ ؟		
37	la douleur est-elle aggravée par quelque chose ?	هَلْ اشْتَدَّ الأَلَمُ لِسَبَبٍ مَا ؟		
38	avez-vous pris ces médicaments dans le passé ?	هَلْ تَنَاوَلْتَ هَذِهِ الأَدْوِيَةِ فِي الْمَاضِي ؟		
39	êtes-vous stressé ?	هَلْ تَشْعُرُ بِالتَّوَتُّرِ ؟		
40	quel est votre métier ?	مَاهِيَ مِهْنَتُك ؟		
41	vous sentez-vous reposé au réveil	هَلْ تَشْعُرُ بِالرَاحَةِ عِنْدَ اسْتِيقَاظِكَ مِنْ النَّوْمِ ؟		
42	la douleur est-elle soulagée par quelque chose ?	هَلْ خَفَّ الأَلَمُ لِسَبَبٍ مَا ؟		
43	avez-vous déjà eu ce type de douleur ?	هَلْ شَعَرْتَ بِهَذَا الأَلَمِ مِنْ قَبْلُ ؟		
44	avez-vous très mal ?	هَلْ تَشْعُرُ بِأَلَمٍ شَدِيدٍ ؟		
45	vomissez-vous ?	هَلْ تَتَقَيَّأ ؟		
46	avez-vous des problèmes de vision ?	هَلْ لَدَيْك مَشَاكِلُ بَصَرِيَّة ؟		
47	avez-vous pris un médicament dans le passé ?	هَلْ تَنَاوَلْتَ دَوَاءً مَا فِي الْمَاضِي ؟		
48	faites-vous de l'exercice physique ?	هَلْ تَقُومُ بِتَمَارِينَ رِيَاضِيَة ؟		
49	avez-vous des difficultés à exercer votre métier ?	هَلْ تَجِدُ صُعُوبَةً فِي مُزَاوَلَةِ مِهْنَتِك ؟		
50	avez-vous du mal à dormir la nuit ?	هَلْ تَجِدُ صُعُوبَةً فِي النَّوْمِ فِي اللَّيْلِ ؟		
51	alaitez-vous ?	هَلْ تُرْضِعِينَ ؟		
52	pouvez-vous me montrer ce qui vous amène ?	مَا هُوَ سَبَبُ قُدُومِكِ إِلَى هُنَا ؟		
53	avez-vous mal quand vous allez à selles ?	هَلْ تَشْعُرِينَ بِأَلَمٍ أَثْنَاءَ التَّبَرُّز ؟		
54	prenez-vous un médicament tous les jours ?	هَلْ تَتَنَاوَلِينَ دَوَاءً مَا كُلَّ يَوْم ؟		
55	c'est la première fois que ça vous arrive ?	هَلْ هَذِهِ أَوَّلُ مَرَّةٍ تَشْعُرِينَ فِيهَا بِهَذَا الأَلَم ؟		
56	avez-vous mal ailleurs ?	هَلْ تَشْعُرِينَ بِأَلَمٍ فِي أَمَاكِنَ أُخْرَى مِنَ الْجِسْمِ ؟		
57	y a-t-il du sang dans les urines ?	هَلْ يَحْتَوِي بَوْلُكِ عَلَى دَم ؟		
58	avez-vous mal dans le bas du dos ?	هَلْ تَشْعُرِينَ بِأَلَمٍ فِي أَسْفَلِ الظَّهْرِ ؟		
59	avez-vous mal au ventre ?	هَلْ تَشْعُرِينَ بِأَلَمٍ فِي الْبَطْنِ ؟		
60	avez-vous vos règles en ce moment ?	هَلْ أَنْتِ فِي دَوْرَتِكِ الشَّهْرِيَّةِ الآنَ ؟		
61	êtes-vous tombé dernièrement ?	هَلْ سَقَطْتَ عَلَى الأَرْضِ مُؤَخَّرًا ؟		
62	avez-vous pris des médicaments contre la douleur dans le passé ?	هَلْ تَنَاوَلْتَ أَدْوِيَةً مُسَكِّنَةً لِلأَلَمِ فِي الْمَاضِي ؟		
63	avez-vous une maladie particulière ?	هَلْ تُعَانِي مِنْ مَرَضٍ مُحَدَّدٍ ؟		
64	avez-vous des problèmes gynécologiques ?	هَلْ تُعَانِينَ مِنْ مَشَاكِلَ نِسَائِيَّةٍ أَوْ إِنْجَابِيَة ؟		
65	quelle est la date de vos dernières règles ?	مَا هُوَ تَارِيخُ دَوْرَتِكِ الشَّهْرِيَّةِ الأَخِيرَة		

Summary	Raw
correct	0
ambigous	0
mistranslation	0
nonsense	0
Total adequacy score	**0**

A.4 Corpus BabelDr: Comprehensibility

Comprehensibility

	Score
fluent	4
non-idiomatic	3
syntax error	2
incomprehensible	1

ID	Translation	Score
1	هَلْ تَشْكِينَ مِنَ الْحَكَّة ؟	
2	هَلِ الأَلَمُ يُشْبِهُ الْحُرْقَة ؟	
3	هَلْ تَتَنَاوَلُ دَوَاءً آخَرَ ؟	
4	هَلْ تَشْعُرِينَ بِأَلَمٍ فِي أَمَاكِنَ أُخْرَى مِنَ الْجِسْمِ ؟	
5	هَلْ تَمَّ تَشْخِيصُ أَيَّةِ حَسَاسِيَّةٍ لَدَيْكَ ؟	
6	هَلْ لَدَيْكِ الْحُمَّى ؟	
7	هَلْ تُلَاحِظُ ظُهُورَ أَعْرَاضٍ مُعَيَّنَةٍ الآنَ ؟	
8	هَلْ يُمْكِنُكِ الإِشَارَةُ بِالأَصْبُعِ إِلَى مِنْطَقَةِ الأَلَمِ ؟	
9	هَلْ تَشْكِينَ مِنْ سُعَالٍ جَافٍ ؟	
10	هَلْ تُعَانِينَ مِنْ مَرَضٍ حَادّ ؟	
11	هَلْ أَنْتِ حَامِل ؟	
12	هَلِ الأَلَمُ شَدِيدٌ ؟	
13	مُنْذُ كَمْ يَوْم ؟	
14	مَرْحَبًا	
15	هَلْ تَشْعُرِينَ بِأَلَمٍ عِنْدَ التَّبَوُّل ؟	
16	مَاهِيَ شِدَّةُ الأَلَمِ عَلَى مِقْيَاسِ دَرَجَاتٍ مِنْ 0 إِلَى 10 عِلْمًا أَنَّ 0 يَعْنِي غِيَابَ الأَلَمِ وَ 10 الْحَدَّ الأَقْصَى لِلأَلَمِ ؟	
17	هَلْ تَنَاوَلْتِ دَوَاءً مَا ؟	
18	هَلْ تَعَرَّضْتِ لِضَرْبَةٍ فِي الرَّأْسِ ؟	
19	مُنْذُ كَمْ يَوْم تَشْعُرُ بِالأَلَمِ ؟	
20	كَمْ عُمْرُكَ ؟	
21	هَلْ تَشْعُرُ بِأَلَمٍ فِي الْكَتِفَيْنِ ؟	
22	هَلْ تَجِدُ صُعُوبَةً فِي شُرْبِ السَّوَائِل ؟	
23	هَلْ لَدَيْكَ مَشَاكِلُ صِحِّيَّةٌ أُخْرَى ؟	
24	هَلْ تَتَنَاوَلُ دَوَاءً مَا ؟	
25	هَلْ تَجِدِينَ صُعُوبَةً فِي الْبَلْع ؟	
26	مُنْذُ كَمْ يَوْم وَأَنْتِ تَسْعُلِينَ ؟	
27	هَلْ تَشْكِينَ مِنْ شَيْءٍ آخَرَ ؟	
28	مُنْذُ كَمْ يَوْم تَشْعُرِينَ بِالأَلَمِ ؟	
29	سَأَتَوَلَّى النَّظَرَ فِي حَالَتِكِ الْيَوْمَ	
30	هَلْ تَنَاوَلْتِ أَدْوِيَةً مُسَكِّنَةً لِلأَلَمِ الْيَوْمَ ؟	
31	هَلْ غَالِبًا مَا تُصَابُ بِالصُّدَاعِ النِّصْفِي ؟	
32	هَلْ تَشْعُرِينَ بِأَلَمٍ فِي مَكَانٍ مَا الآن ؟	
33	هَلْ تُسْعِلِينَ ؟	
34	هَلْ قِسْتَ ضَغْطَ دَمِكَ ؟	
35	هَلْ لَدَيْكَ حَسَاسِيَّةٌ مِنْ بَعْضِ الأَدْوِيَة ؟	
36	هَلْ تَنَاوَلْتَ دَوَاءً مَا الْيَوْمَ ؟	
37	هَلْ اشْتَدَّ الأَلَمُ لِسَبَبٍ مَا ؟	
38	هَلْ تَنَاوَلْتَ هَذِهِ الأَدْوِيَةَ فِي الْمَاضِي ؟	
39	هَلْ تَشْعُرُ بِالتَّوَتُّرِ ؟	
40	مَاهِيَ مِهْنَتُكَ ؟	
41	هَلْ تَشْعُرُ بِالرَّاحَةِ عِنْدَ اسْتِيقَاظِكَ مِنَ النَّوْمِ ؟	
42	هَلْ خَفَّ الأَلَمُ لِسَبَبٍ مَا ؟	
43	هَلْ شَعَرْتَ بِهَذَا الأَلَمِ مِنْ قَبْلُ ؟	
44	هَلْ تَشْعُرُ بِأَلَمٍ شَدِيدٍ ؟	
45	هَلْ تَتَقَيَّأُ ؟	
46	هَلْ لَدَيْكَ مَشَاكِلُ بَصَرِيَّة ؟	
47	هَلْ تَنَاوَلْتَ دَوَاءً مَا فِي الْمَاضِي ؟	
48	هَلْ تَقُومُ بِتَمَارِينَ رِيَاضِيَة ؟	
49	هَلْ تَجِدُ صُعُوبَةً فِي مُزَاوَلَةِ مِهْنَتِكَ ؟	
50	هَلْ تَجِدُ صُعُوبَةً فِي النَّوْمِ فِي اللَّيْلِ ؟	
51	هَلْ تُرْضِعِينَ ؟	
52	مَا هُوَ سَبَبُ قُدُومِكَ إِلَى هُنَا ؟	
53	هَلْ تَشْعُرِينَ بِأَلَمٍ أَثْنَاءَ التَّبَرُّز ؟	
54	هَلْ تَتَنَاوَلِينَ دَوَاءً مَا كُلَّ يَوْم ؟	
55	هَلْ هَذِهِ أَوَّلُ مَرَّةٍ تَشْعُرِينَ فِيهَا بِهَذَا الأَلَم ؟	
56	هَلْ تَشْعُرِينَ بِأَلَمٍ فِي أَمَاكِنَ أُخْرَى مِنَ الْجِسْمِ ؟	
57	هَلْ يَحْتَوِي بَوْلُكِ عَلَى دَم ؟	
58	هَلْ تَشْعُرِينَ بِأَلَمٍ فِي أَسْفَلِ الظَّهْرِ ؟	
59	هَلْ تَشْعُرِينَ بِأَلَمٍ فِي الْبَطْن ؟	
60	هَلْ أَنْتِ فِي دَوْرَتِكِ الشَّهْرِيَّةِ الآنَ ؟	
61	هَلْ سَقَطْتَ عَلَى الأَرْضِ مُؤَخَّرًا ؟	
62	هَلْ تَنَاوَلْتَ أَدْوِيَةً مُسَكِّنَةً لِلأَلَمِ فِي الْمَاضِي ؟	
63	هَلْ تُعَانِي مِنْ مَرَضٍ مُعَدٍّ ؟	
64	هَلْ تُعَانِينَ مِنْ مَشَاكِلَ تَنَاسُلِيَّةٍ أَوْ إِنْجَابِيَة ؟	
65	مَا هُوَ تَارِيخُ دَوْرَتِكِ الشَّهْرِيَّةِ الأَخِيرَة	

Summary	Raw
fluent	0
non-idiomatic	0
syntax error	0
incomprehensible	0
Total comprehensibility score	**0**

A.5 Corpus Google Translate: Adequacy

Adequacy

	Score
correct	4
ambigous	3
mistranslation	2
nonsense	1

ID	Source	Translation	Score	Dangerous
1	Bonjour	صباح الخير		
2	vous avez mal	انت مجروح		
3	êtes-vous enceinte	هل انت حامل		
4	vous avez mal quand vous faites pipi	يؤلمك عندما تتبول		
5	depuis combien de temps	منذ متى		
6	depuis 1 jours	منذ يوم 1		
7	vous avez déjà pris des médicaments	هل سبق لك أن تناولت دواء		
8	est-ce que vous avez du mal à avaler	هل تعانين من صعوبة في البلع		
9	est-ce que vous arrivez à avaler	هل تستطيع أن تبتلع		
10	est-ce que vous avez de la fièvre	هل لديك حمى		
11	est-ce que vous avez des pertes blanches	هل لديك إفرازات بيضاء		
12	donc ce sont des douleurs de gorge	لذلك هوالتهاب في الحلق		
13	où avez-vous mal	أين تشعر بالألم		
14	est-ce que la douleur est forte	هل الألم قوي		
15	avez-vous avez-vous pris des médicaments pour ça	هل تناولت أي دواء لذلك		
16	est-ce que c'est la première fois	هل هذه هي المرة الأولى		
17	vous avez déjà eu mal comme ça avant	لقد عانيت من ألم مثل هذا من قبل		
18	est-ce que vous avez d'autres symptômes	هل لديك أي أعراض أخرى		
19	prenez-vous des médicaments quotidiens	هل تتناول أدوية يومية		
20	est-ce que ça gratte	هل تخدش		
21	vous aviez pris quoi la dernière fois	ماذا اخذت آخر مرة		
22	est-ce que ça vous arrive souvent pendant la période d'hiver	هل يحدث هذا لك غالبًا خلال فترة الشتاء		
23	est-ce que c'est lié à quelque chose en particulier un événement	هل هو متعلق بشيء معين يحدث معين		
24	est-ce que vous avez une idée comment le soigner	هل لديك أي فكرة عن كيفية علاجه		
25	est-ce que vous avez mal en allant aux toilettes	هل تشعرين بألم عند الذهاب إلى المرحاض		
26	est-ce que vous avez mal dans le bas du dos	هل تعانين من الم في اسفل ظهرك		
27	est-ce que vous avez de la fièvre	هل لديك حمى		
28	avez-vous de la toux	هل لديك سعال		
29	est-ce que ça brule quand vous allez aux toilettes	هل تحترق عندما تذهب إلى المرحاض		
30	qu'est-ce qui vous amène aujourd'hui	ما الذي أتى بك اليوم		
31	avez-vous mal quelque part	هل تؤذي في مكان ما		
32	depuis combien de temps avez-vous mal à la tête	منذ متى وأنت تعاني من الصداع		
33	avez-vous mal aussi autre part	هل تشعر بألم في مكان آخر		
34	avez-vous d'autres symptômes	هل لديك أي أعراض أخرى		
35	avez-vous le nez bouché	هل لديك انسداد في الأنف		
36	est-ce que vous avez un bébé	هل لديك طفل		
37	avez-vous pris des médicaments pour votre maux de tête	هل لديك سعال		
38	prenez-vous d'autres médicaments habituellement	هل عادة ما تتناول أي أدوية أخرى		
39	avez-vous des problèmes de santé	هل لديك أي مشاكل صحية		
40	voulez-vous un médicament pour le mal de tête	هل تريد دواء للصداع		
41	c'est la première fois que ça vous arrive	هذه هي المرة الأولى التي يحدث فيها هذا لك		
42	ça vous arrive souvent	يحدث لك في كثير من الأحيان		
43	c'est quand la dernière fois que ça vous est arrivé	متى كانت آخر مرة حدث فيها هذا لك		
44	où est-ce que vous avez mal	اين تؤذي		
45	où sont les autres symptômes	أين الأعراض الأخرى		
46	depuis combien de jours	كم يوما		
47	est-ce que vous avez plus mal à droite ou à gauche	هل تشعر بألم أكبر في اليمين أو اليسار		
48	est-ce que vous avez mal au niveau des yeux	هل تعانين من الم في عينيك		
49	vous avez des douleurs ailleurs	لديك ألم في مكان آخر		
50	vous pouvez me montrer où vous avez mal	يمكنك أن تريني أين تؤذي		
51	est-ce que c'est la même douleur toute la journée	هل هو نفس الألم طوال اليوم		
52	est-ce que vous avez déjà une douleur comme ça avant	هل عانيت من ألم مثل هذا من قبل		
53	est-ce que vous avez des pertes vaginales	هل تعانين من إفرازات مهبلية		
54	est-ce que vous avez déjà pris les médicaments pour votre mal de tête	هل سبق لك تناول دواء للصداع		
55	est-ce que vous avez des maladies chroniques ou d'autres problèmes de santé	هل لديك أي أمراض مزمنة أومشاكل صحية أخرى		
56	est-ce que vous êtes enceinte où vous êtes en train d'allaiter	هل انت حامل او مرضعة		
57	est-ce que vous avez mal autre part	هل تشعر بألم في مكان آخر		
58	depuis combien de jours vous avez mal à la tête	كم يوم عانيت من صداع		
59	est-ce que c'est la première fois que ça vous arrive?	هل هذه هي المرة الأولى التي يحدث فيها هذا لك؟		
60	est-ce que vous avez déjà pris un médicament pour le mal de tête	هل سبق لك تناول دواء للصداع		
61	est-ce que vous prenez des autres médicaments tous les jours	هل تتناول أي أدوية أخرى كل يوم		
62	est-ce que vous êtes enceinte	هل انت حامل		
63	ça vous fait mal quand vous faites pipi	يؤلمك عندما تتبول		
64	sur une échelle de 1 à 10 vous avez mal combien	على مقياس من 1 إلى 10 مدى الألم الذي تشعر به		
65	vous avez très mal	أنت تتألم بشدة		
66	depuis combien de temps vous avez mal	منذ متى وأنت تتألم		
67	est-ce que vous avez de la fièvre	هل لديك حمى		
68	est-ce que c'est rouge	هل هي حمراء		
69	est-ce que vous avez de la toux	هل لديك سعال		
70	avez-vous d'autres symptômes	هل لديك أي أعراض أخرى		
71	avez-vous des vomissements	هل تتقيأ		
72	avez-vous mal à la tête	هل تعاني من صداع في الرأس		
73	avez-vous mal autre part	هل تشعر بألم في مكان آخر		
74	pouvez-vous me montrer où vous avez mal	هل يمكنك أن تريني أين تؤذي		
75	avez-vous déjà pris des médicaments	هل سبق لك أن تناولت أي دواء		
76	avez-vous une maladie chronique	هل لديك مرض مزمن		
77	est-ce que ça arrive tous les mois	هل هذا يحدث كل شهر		
78	avez-vous de la diarrhée	هل لديك اسهال		
79	bonjour où avez-vous mal	مرحبا اين تؤذي		
80	vous avez mal quand vous avalez	يؤلمك عندما تبتلع		
81	est-ce que vous avez du sang quand vous faites pipi	هل لديك دم عند التبول		
82	avez-vous pris des médicaments	هل تناولت أي دواء		
83	pas de fièvre	لا حمى		
84	Bonjour Madame	صباح الخير سيدتي		
85	où elle est localisé la douleur	حيث يقع الألم		
86	quel âge avez-vous	كم عمرك		
87	est-ce que vous pouvez m'indiquer la partie du corps qui vous dérange	هل يمكن أن تخبرني أي جزء من الجسم يزعجك		
88	dans quelle partie du corps ils sont situés ces symptomes	في أي جزء من الجسم توجد هذه الأعراض		
89	est-ce que vous prenez des médicaments	هل تتعاطى أي أدوية		
90	est-ce que vous avez envie de vomir ou la tête qui tourne	هل تشعر بالرغبة في التقيؤ أو الدوار		
91	est-ce que c'est au niveau génital	هل هو على مستوى الأعضاء التناسلية		
92	avez-vous de la fièvre	هل لديك حمى		
93	avez-vous des maladies chroniques	هل لديك أي أمراض مزمنة		
94	ça vous gratte la gorge	إنه يخدش حلقك		
95	est-ce que les ganglions sont gonflés	هي الغدد الليمفاوية منتفخة		
96	vous pouvez me dire où vous avez mal	يمكنك أن تخبرني أين آذيت 1 كيلو		

Summary	Raw
correct	0
ambigous	0
mistranslation	0
nonsense	0
Total adequacy score	**0**

A.6 Corpus Google Translate: Comprehensibility

Comprehensibility

	Score
fluent	4
non-idiomatic	3
syntax error	2
incomprehensible	1

ID	Translation	Score
1	صباح الخير	
2	انت مجروح	
3	هل انت حامل	
4	يؤلمك عندما تتبول	
5	منذ متى	
6	منذ يوم 1	
7	هل سبق لك أن تناولت دواء	
8	هل تعانين من صعوبة في البلع	
9	هل تستطيع أن تبتلع	
10	هل لديك حمى	
11	هل لديك إفرازات بيضاء	
12	لذلك هو التهاب في الحلق	
13	أين تشعر بالألم	
14	هل الألم قوي	
15	هل تناولت أي دواء لذلك	
16	هل هذه هي المرة الأولى	
17	لقد عانيت من ألم مثل هذا من قبل	
18	هل لديك أي أعراض أخرى	
19	هل تتناول أدوية يومية	
20	هل تخدش	
21	ماذا اخذت آخر مرة	
22	هل يحدث هذا لك غالبًا خلال فترة الشتاء	
23	هل هو متعلق بشيء معين يحدث معين	
24	هل لديك أي فكرة عن كيفية علاجه	
25	هل تشعرين بألم عند الذهاب إلى المرحاض	
26	هل تعانين من الم في اسفل ظهرك	
27	هل لديك حمى	
28	هل لديك سعال	
29	هل تحترق عندما تذهب إلى المرحاض	
30	ما الذي آلى بك اليوم	
31	هل تؤذي في مكان ما	
32	منذ متى وأنت تعاني من الصداع	
33	هل تشعر بألم في مكان آخر	
34	هل لديك أي أعراض أخرى	
35	هل لديك انسداد في الأنف	
36	هل لديك طفل	
37	هل لديك سعال	
38	هل عادة ما تتناول أي أدوية أخرى	
39	هل لديك أي مشاكل صحية	
40	هل تريد دواء للصداع	
41	هذه هي المرة الأولى التي يحدث فيها هذا لك	
42	يحدث لك في كثير من الأحيان	
43	متى كانت آخر مرة حدث فيها هذا لك	
44	اين تؤذي	
45	أين الأعراض الأخرى	
46	كم يوما	
47	هل تشعر بألم أكبر في اليمين أو اليسار	
48	هل تعانين من الم في عينيك	
49	لديك ألم في مكان آخر	
50	يمكنك أن تريني أين تؤذي	
51	هل هو نفس الألم طوال اليوم	
52	هل عانيت من ألم مثل هذا من قبل	
53	هل تعانين من إفرازات مهبلية	
54	هل سبق لك تناول دواء للصداع	
55	هل لديك أي أمراض مزمنة أو مشاكل صحية أخرى	
56	هل انت حامل او مرضعة	
57	هل تشعر بألم في مكان آخر	
58	كم يوم عانيت من صداع	
59	هل هذه هي المرة الأولى التي يحدث فيها هذا لك؟	
60	هل سبق لك تناول دواء للصداع	
61	هل تتناول أي أدوية أخرى كل يوم	
62	هل انت حامل	
63	يؤلمك عندما تتبول	
64	على مقياس من 1 إلى 10 مدى الألم الذي تشعر به	
65	أنت تتألم بشدة	
66	منذ متى وأنت تتألم	
67	هل لديك حمى	
68	هل هي حمراء	
69	هل لديك سعال	
70	هل لديك أي أعراض أخرى	
71	هل تتقيأ	
72	هل تعاني من صداع في الرأس	
73	هل تشعر بألم في مكان آخر	
74	هل يمكنك أن تريني أين تؤذي	
75	هل سبق لك أن تناولت أي دواء	
76	هل لديك مرض مزمن	
77	هل هذا يحدث كل شهر	
78	هل لديك اسهال	
79	مرحبا اين تؤذي	
80	يؤلمك عندما تبتلع	
81	هل لديك دم عند التبول	
82	هل تناولت أي دواء	
83	لا حمى	
84	صباح الخير سيدتي	
85	حيث يقع الألم	
86	كم عمرك	
87	هل يمكن أن تخبرني أي جزء من الجسم يزعجك	
88	في أي جزء من الجسم توجد هذه الأعراض	
89	هل تتعاطى أي أدوية	
90	هل تشعر بالرغبة في التقيؤ أو الدوار	
91	هل هو على مستوى الأعضاء التناسلية	
92	هل لديك حمى	
93	هل لديك أي أمراض مزمنة	
94	إنه يخدش حلقك	
95	هي الغدد الليمفاوية منتفخة	
96	يمكنك أن تخبرني أين آذيت 1 كيلو	

Summary

	Raw
fluent	0
non-idiomatic	0
syntax error	0
incomprehensible	0
Total comprehensibility score	**0**