# 2

# Speech and Translation Technologies

## *Healthcare Applications*

MARK SELIGMAN

## 2.1 Introduction

Cross-language communication in healthcare is urgently needed. Daily and nightly throughout the world, thousands of conversations are required between caregivers – doctors, nurses, administrators, volunteers, and others – and patients or family members with differing native languages.

Chapter 1 describes and illustrates the exploding development of the relevant linguistic technologies – machine translation (MT) of text, automatic speech recognition (ASR), and text-to-speech (TTS). The related infrastructure – wireless communication, cloud computing, and mobile devices – has also been developing apace. This chapter will shift focus to the combination and application of these technologies in the healthcare context, with special interest in speech translation.

Given this impressive and accelerating progress, we'd expect various automatic translation and speech-enabled systems to be in widespread use by now; in fact, however, adoption remains slow. We'll examine the obstacles to adoption and directions for overcoming them in Section 2.2. In Section 2.3, we'll examine two major types of speech translation systems, concentrating on their respective approaches to the same obstacles. Section 2.4 will survey some healthcare-oriented communication systems, past and future. We'll conclude with an optimistic forecast for speech and translation applications in healthcare, tempered by due cautions.

## 2.2 Obstacles to Adoption and Potential Solutions

One key factor in the lagging adoption of linguistic technology in healthcare is the sheer difficulty of understanding the relevant technologies, and thus the natural hesitation to trust them. Accordingly, Chapter 1 aimed to bridge the



https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

understanding gap for healthcare workers by explaining speech recognition, speech synthesis, and MT.

However, even if potential users of speech and translation technologies in the healthcare field can gain sufficient understanding to realistically evaluate specific implementations, obstacles will remain. It will be helpful to group these under two major headings: reliability and customization per use case.

### 2.2.1 Reliability

In any field with demanding communication requirements, workers will hesitate to employ even exciting and progressing communication technology if they fear it may cause embarrassing, or even dangerous, errors. This tendency is compounded in the healthcare field, where communication errors can indeed have disastrous consequences. And it is further compounded for translation technology in particular, since users have until now usually been unable to judge the correctness of the results and have been unable to correct any errors even if recognized. Measurable *accuracy* is increasing in the three technologies of interest – most dramatically for translation, as informally demonstrated in Chapter 1; but this progress alone is unlikely to overcome high-tech hesitancy. Reliability must also be measured in terms of potential users' *confidence* – a psychological rather than technical matter. For our purposes, then, "reliability" implies accuracy *plus* trustworthiness. Trust can be fostered in several ways.

#### 2.2.1.1 Offline Preparation of Output

Trust can be maximized through use of professionally prepared or confirmed output, as opposed to output generated on the spot. And, in fact, pre-vetted output of text translation is one important element of the approach to speech translation taken by fixed-phrase-based speech translators like BabelDr (Spechbach and Bouillon, 2019; Chapter 5). Because translations are prepared in advance by professionals, they can be assumed trustworthy – at least to the extent that one can trust the processes that select the appropriate prechecked translations by matching them against source-language inputs to be translated. (The matching processes are discussed in Chapter 1.)

Even in systems offering full translation, professionally prepared translations (or translations previously confirmed through other means to be discussed later) can be used in a preliminary step: If a sufficiently close match to the current input is found in the database of stored translations, the match's prepared translation will be used; if not, the input is passed to the subsystem designed for full translation. The repository of prepared translations thus serves

as *translation memory*, an element of many MT systems, whether rule-based, statistical, or neural. This approach was employed, for instance, in the Converser for Healthcare prototype speech translation system (Seligman and Dillinger, 2006a, 2006b, 2008, 2011, 2012, 2013, 2014, 2015, 2016; Dillinger and Seligman, 2004; Zong and Seligman, 2005) under the proprietary name of Translation Shortcuts™. Section 2.3.3.7 will demonstrate its use in context.

Prepared translations can also be accessed more directly by literate users through text listings of the phrases to be translated, which can be browsed or automatically searched. To facilitate browsing or search, the listings can be categorized: For example, prepared translations can be categorized for pharmacy, nursing, or eye-care use; and translations for pharmacy can be subcategorized for consultation, prescription pickup, and so on (Section 2.3.2.7).

#### 2.2.1.2 Feedback

Another approach to fostering trust is to provide effective feedback: Rather than blindly trusting speech recognition and translation outputs, users can see or hear recognition results and native-language retranslation, and perhaps correct any errors. For speech recognition, literate users can profit from textual feedback; and to enable eyes-free use or for illiterates, playback via TTS could additionally be offered. For MT, *back-translation* – that is, translation from the target language back to the original source language – can help to check whether a preliminary translation has conveyed the intended meaning. Various techniques can be applied to minimize back-translation errors (Seligman and Dillinger, 2014). Back-translation has usually been given only textually, but auditory feedback via TTS is also possible.

#### 2.2.1.3 Correction

If users can be enabled to *recognize* errors, it may be feasible to enable error *correction* as well. For speech recognition, assuming results are made visible in text, literate users can correct any errors by first selecting the erroneous segment and then manually entering or pronouncing the correction.

With respect to translation errors, correction by monolingual users is more challenging, but still possible.

- A "Proceed with Caution Mode" can be offered, in which a preliminary back-translation, monitored by the staff member only, must be approved before transmission to the patient is authorized. If an error is seen, the users' paraphrase of the input, or of a selected part of it, may lead to a translation that can be approved.

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

- Users may note specific ambiguity errors in a back-translation – indicating translation of the wrong meaning for words or expressions with multiple meanings – such as translation of English *cool* as "chilly, nippy, somewhat cold . . . " when "awesome, terrific, fantastic . . . " was intended. They can then be enabled to select the erroneous segment and to choose among alternative meanings, which can be indicated in the native language via synonyms, definitions, examples, or pictures.

  These correction possibilities are illustrated in context in Section 2.3.2.7.

**Too Much Trouble?** Monitoring of speech recognition and translation results inevitably takes attention and time, and any correction even more so. However, depending on the use case, the benefits in real-time accuracy and trust may sometimes justify the effort. Again, in healthcare, disastrous translations must be avoided at all costs.

And there is another justification for taking the trouble to correct, when and if enabled: Corrections can be captured and used in several ways. First, the corrections can become training material for machine learning that can substantially improve the systems in question. Corrections can be domain-specific, so that training of speech recognition and translation can be optimized for specific use cases. Second, corrected translations can be considered to have passed the trust test, and thus to have qualified as entries in translation memory.

Useful or not, correction mechanisms will likely be resented as intrusive unless interface facilities are provided for turning them on or off as appropriate. One system employed icons allowing switching between "Full Speed Ahead" mode, in which no verification stage would be used, and "Proceed with Caution" mode, in which a pause for verification would be imposed.

- Earring icons controlled handling of speech recognition: Selection of a green earring meant that ASR results would be sent to translation immediately, without pausing for pre-checking, while choice of a yellow icon did impose a pause. A red earring stopped all speech recognition, to block accidental use.
- A Traffic Light Icon controlled handling of translation: Green meant that translations would be immediately transmitted to users, while yellow meant that a verification dialogue would be presented first. A red light stopped all translation, to prevent accidental use.

  These interface facilities, too, are illustrated in context in Section 2.3.2.7.

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

#### 2.2.1.4 Record-Keeping

A final approach to building trust is regular recording of conversations. While audio recordings would be possible, transcripts may be enough for most purposes. These should include both the original inputs and the automatic translations. It may also be helpful to include any back-translations, so that monolingual staff or researchers can post-verify communication.

### 2.2.2 Customization per Use Case

One way to overcome obstacles to widespread use of speech and translation technologies within the healthcare field, we've suggested, is to ensure reliability – again, entailing not only increased measurable accuracy but increased user confidence. Another way is to ensure that the technologies can be used conveniently and practically in each use case – in other words, to ensure customization of the technologies per use case. Our motto here: "Magic is not enough!" While the relevant technologies have achieved levels of performance that would have seemed miraculous at the turn of the millennium, we've learned that awe alone cannot bring them across the proverbial chasm toward general acceptance. In every demanding field, but especially in healthcare, responsible people are overloaded and properly conservative. The tools must be not only trustworthy but transparently easy to use and seamlessly convenient: They must fit the individual use cases like gloves. Fitting those gloves takes time and financial support, so implementing a solution becomes an organizational and business issue.

#### 2.2.2.1 Platforms

The devices and software required for delivery of speech and translation services have evolved quickly. To dramatize the difference a decade makes, here's a look back at the equipment used in 2011 for a three-month pilot project involving full speech translation at a San Francisco hospital (Seligman and Dillinger, 2015).

At that time, most of the infrastructure we now take for granted was in the future:

- There were no modern flat tablets, so thick and heavy portable devices with built-in handles were used. An alternative setup aimed to accommodate staff members and patients facing each other across a desk. Staff could operate the full interface on a desktop computer serving as master, while patients could see, but not manipulate, a secondary computer showing the same view. For both arrangements, setup and maintenance were time-consuming and error prone.

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

- iPhones had appeared in 2007, but on-device memory capacity was limited, so processing was strained when running full speech translation. (Jibbigo (Eck et al., 2010) – a spinoff of Carnegie Mellon University research under Alex Waibel, later sold to Facebook – nevertheless released several on-device systems for individual language pairs, but their reliability was insufficient for demanding use cases.)
- Remote computing was thus a tempting alternative, but cloud computing was immature, and locally installed software remained the only practical option, with the attendant installation and maintenance headaches.
- Speech recognition software was still speaker-dependent, so each user needed to provide a voice sample during a short training session – doable for staff members, despite some scheduling annoyances, but impractical for patients, so translation was restricted to direction. (Soon after the pilot, zero-training ASR arrived, so that voice input from both sides would have been enabled.)
- Web conferencing with video was nascent and awkward to arrange, so extensive software integration work would have been required to enable remote conferencing with automatic translation.

Thankfully, ten years on, these handicaps have now been alleviated or resolved:

- A wide selection is now available of light and powerful computing devices – smartphones and tablets of various sizes.
- Cloud computing is now standard, making software installation a trivial matter of app download and registration.
- Speaker-independent speech recognition requiring no preparatory training is now taken for granted.
- Web conferencing has almost overnight become universal – with a boost from the pandemic era – and movement toward multilingual meeting capability is well underway. (Zoom has recently acquired relevant software (Marking, 2021).)

However, it is still proving difficult to engineer speech translation systems that offer an acceptable combination of reliability and use case customization. Ergonomic design is particularly challenging. An ideal system would be as unobtrusive as a skilled interpreter: It would offer highly reliable translation while allowing completely hands-free and eyes-free operation.

Unfortunately, two of the three major components of a speech translation system for demanding use cases like healthcare – speech recognition for spontaneous speech in noisy environments and automatic full translation – will still

require some user monitoring until accuracy reaches an extremely high threshold, and may well continue to do so even beyond that point for user confidence ("reliability"). But that monitoring must be enabled without undue distraction from the work at hand.

Standard smartphones, tablets, and laptops are now available, but each format has its plusses and minuses. Smartphones, for instance, are easily portable, but their screens are small, so feedback may be hard to read at a distance; phones' onboard memory capacities are constrained; their speakers are limited in volume; their microphones may not work well if the device is far from a speaker; and, unless a holder is used, at least one hand will be occupied. Comparable pros and cons will apply to tablets and laptops.

In view of these issues, attempts have been made to build dedicated devices specialized for speech translation, and even for healthcare specifically. Fujitsu's Artificial Intelligence Laboratory, for instance, mounted a project to create two specialized microphone formats for this purpose. This effort is fully discussed in Section 2.3.2.9.

Several other companies have undertaken development of comparable dedicated devices for speech translation. Presently on sale are portable or wearable items marketed as ili, Pocketalk, Cheetah, and others. Voice translation is also available for Apple Watch, as powered by iTranslate, IHG Translator App, Speak & Translate, Microsoft Translator, Babbel, and TalkMondo. All these offerings address ergonomic – and therefore customization – issues to some degree, but none yet support reliability facilities.

In choosing between such dedicated devices and more standard ones like mobile phones and tablets, tradeoffs are unavoidable. Specialized equipment may be more narrowly directed at the given use case – desirable in principle, as we've stressed. But of course standard devices are everywhere: inexpensive, easy to obtain, familiar to use, and home to many apps developed at the makers' expense. In the end, the tradeoffs may fade as standard devices and software become ever more capable and versatile.

#### 2.2.2.2 Peripherals

For any of the speech translation platforms just surveyed, auxiliary peripheral devices could provide ergonomic enhancements, some of which might prove decisive for usability per use case. For example, while auditory feedback for staff could boost reliability for both speech recognition and translation, it's likely to be confusing for patients. Earbuds – now connectable via Bluetooth to most devices – could ensure that only staff members heard appropriate confirmations.

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

More generally, Augmented Reality (AR) is quickly gaining popularity, and will likely experience an explosion with the imminent arrival of smart glasses. These will allow visual feedback to appear as "heads-up" displays visible through the lenses without head movement, and to be heard through embedded your-ears-only speakers. Translations, too, will be viewable and audible in the same way for both staff and patients.

Some AR devices will support not only visualization and sound for visual and auditory feedback, but also control capabilities for correction and guidance. They'll track hand movements in relation to virtual displays – if not immediately and affordably, then later. We can then expect AR to finally enable creation of maximally hands-free and eyes-free interfaces, and in this way to finally combine customization and reliability effectively.

#### 2.2.2.3 Security

Translation and speech programs can be designed to run entirely in the "cloud" – that is, on servers which communicate with devices like smartphones or tablets; to run entirely on those devices; or to run in hybrid modes, with some elements (e.g., translation) online, and some (e.g., speech recognition and TTS) on the device. Related architecture decisions depend on the programs' processing requirements, on the necessary response time, and so on.

Most healthcare organizations worry about data security – certainly to protect their own operations, but often also to meet governmental requirements, such as those of the Health Insurance Portability and Accountability Act (HIPAA) in the US. Patient healthcare information is especially sensitive. Security requirements inevitably complicate adoption of technology for translation and speech. If associated software runs online (in the cloud), many organizations require that it be hosted on their own, usually local, servers. If the software runs on the device, it must often be integrated in approved official software *builds* (program sets).

## 2.3 Speech Translation Designs for Healthcare

Having discussed obstacles to adoption of speech translation systems, broadly grouped as relating to reliability and customization, and having considered a range of current and potential solutions, we now turn to examination of speech translation systems themselves.

Efforts to combine speech and translation technologies for healthcare have sorted themselves into two clear categories, largely based on systems' approaches to the tension between two major goals: On one hand, reliability

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

is paramount in healthcare, as discussed, but wide applicability is also desirable. A tradeoff between these objectives is inevitable, since increased range will always give the opportunity for more errors. Several speech translation systems aiming for maximum reliability have opted for phrase-based design, while those aiming for greater range have risked full (that is, wide-ranging or relatively unrestricted) translation, while sometimes including phrase-based components.

Compromises between the phrase-based and full translation approaches are also possible, as already mentioned. A system enabling full translation can include a preliminary phrase-based stage, in which the input is matched against the set of remembered phrases, already supplied with prepared translations. We have introduced this approach as that of translation memory (Section 2.2.1.1). And again, a system can allow full translation when appropriate, while attempting to mitigate the associated risk of errors through facilities for verification and correction. Corrected translations can then enter the list of pretranslated inputs – that is, they can enter translation memory (Section 2.2.1.3.1). (Both of these strategies are illustrated in Section 2.3.2.7.)

We'll now look at phrase-based and full speech translation systems in turn.

### 2.3.1 Phrase-Based Speech Translation for Healthcare

Several healthcare-oriented systems have been designed to handle pretranslated phrases only, rather than attempting to provide full MT of wide-ranging input. This design decision addresses both of our desiderata: It enhances reliability because it depends on (usually professional) translation in advance, and it aids customization per use case in that relevant phrases can be brought into the system as needed. Here, we'll look at three strictly phrase-based systems.

#### 2.3.1.1 S-MINDS and Phraselator

An early healthcare entry was the S-MINDS system by Sehda, Inc. (later Fluential) (Ehsani et al., 2008). At its center was an extensive set of fixed and pretranslated phrases, and the task of speech recognition was to match the appropriate one so as to enable pronunciation of its translation via TTS. A proprietary facility yielded the best available fuzzy match when no precise match was found. In this respect, the system represented further development of speech translation systems like the earlier Phraselator,[1] a ruggedized

[1] "Phraselator." *Wikipedia*, Wikimedia Foundation, 4 December 2021, at 20: 52(UTC), https://en.wikipedia.org/wiki/Phraselator.

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

handheld device likewise offering translation of fixed phrases only, provided in large quantities to the US military for use in the first Gulf War and later in various military, law enforcement, and humanitarian operations. To provide more flexible speech input, later versions of the Phraselator added technology licensed from Jibbigo (Eck et al., 2010), a commercial system for full speech translation produced by the research group of Alex Waibel.

#### 2.3.1.2 BabelDr

The BabelDr system (Spechbach and Bouillon, 2019; Chapter 5), implemented by a team at the University of Geneva, is a recent example of phrase-only speech translation for healthcare. The system imposes two further constraints: (1) those phrases should generally be yes/no questions and (2) translation is unidirectional, in that patients are expected to respond only nonverbally to translated questions from healthcare staff. As compensation, however, these limitations enhance the overall practicality of the system by reducing the opportunities for error and the need for training. In addition, several interface refinements increase system reliability and facilitate customization for various use cases.

In terms of reliability, the system features transformation of each spoken or typed input phrase into a canonical text phrase, for which a translation has already been supplied and is ready for immediate transmission. In this respect, the system is comparable to the Phraselator and Sehda/Fluential speech-translation systems, as already described.

Importantly, however, the canonical phrases can also provide feedback to users concerning translation accuracy. This useful verification source has not previously been exploited. Experiments supplying confirmatory back-translations via neural networks appear promising as well (Mutal et al., 2019; DeepL commercial translation system[2]). For comparison, other feedback sources which have been used to date include semantically controlled back-translation (Seligman and Dillinger, 2016) and paraphrases generated via interlingua-based semantic representations (Gao et al., 2006).

Regarding customizability: In addition to using canonical phrases for verification of translations, users can access them more directly by browsing or by searching via keywords. The associated translations can then be transmitted without the need for further checking. Users can also focus on desired phrases by indicating the relevant topic through the GUI. These facilities enable quick

[2] "DeepL Translator." *Wikipedia*, Wikimedia Foundation, 10 August 2022, at 17: 37(UTC), https://en.wikipedia.org/wiki/DeepL_Translator.

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

customization of the system since new sets of canonical phrases and their translations can be created quickly.

While several central elements of BabelDr – robust matching of ASR results against a canonical set of pretranslated phrases, feedback to users concerning translation accuracy, enablement of searching and browsing among the phrase set – have been introduced by previous systems, this combination is new and promises to be especially practical, thanks to the imposed limitations and to the innovative handling of feedback.

Due in part to the same limitations, the reported evaluations demonstrate convincingly the usability of the system for successful diagnosis of simulated patients. Also reported are interesting results concerning the relative usability of speech as compared to text input.

### 2.3.2 Full Speech-to-Speech Translation for Healthcare

But now, on to full speech translation, in which vocabulary and grammar is relatively unrestricted – "relatively" because systems may still differ in the expected range of topics: Some may expect (and be trained on) only pharmaceutical matters, for instance, while others may invite conversations on roughly any topic.

Following decades of anticipation, automatic spoken-language translation (SLT) has finally entered widespread use. The Google Translate application, for instance, can bridge dozens of languages in face-to-face conversations, switching languages automatically. Microsoft speech-translation software now powers translated video chat among thirty languages, with sophisticated measures for cleaning up the stutters, errors, and repetitions of spontaneous speech.

Still emerging, however, are speech-translation systems directed at various demanding and socially significant use cases. Viewed from a high level, the main obstacle to widespread adoption has been that the essential components – speech-recognition and -translation technologies – are still error prone. While the error rates may be tolerable when the technologies are used separately, the errors combine and even compound when used together. The resulting translation output is often below the threshold of usability when accuracy is essential. Consequently, until now, use has been largely restricted to use cases – social networking, travel – in which no representation concerning accuracy is demanded or given.

Not that attempts to field systems for more demanding speech-translation applications have been missing. The Defense Advanced Research Programs Agency of the United States, for instance, has an extensive record of innovative work relating to law enforcement, disaster relief, and translation of

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

broadcast news (Seligman and Waibel, 2019) – and healthcare, our main interest here.[3]

In examining healthcare-oriented SLT systems supporting full translation, we'll take as an example Converser for Healthcare, a prototype for communication between English-speaking healthcare staff and Spanish-speaking patients. The discussion is partly of historical interest since the system's pilot project took place in 2011 – an eon ago in computer years; however, most of the issues raised by that evaluation remain current. The system is also handy for present illustrative purposes, since it incorporated in a single application most of the reliability and customization features discussed: It applied interactive verification and correction techniques to the problem of reliability and offered Translation Shortcuts™, a form of translation memory, as its main aid to customization per use case.

Presently, we'll also touch on Fujitsu's healthcare-oriented system, emphasizing ergonomics as its customization approach (Section 2.3.2.2).

#### 2.3.2.1 Converser for Healthcare

Converser was specialized for the healthcare market since the demand was most evident there. At the time of the pilot project, for example, San Francisco General Hospital received more than 3,500 requests for interpretation per month, or 42,000 per year, for 35 different languages. Requests for medical interpretation services are distributed among many wards and clinics (Paras et al., 2002). The resulting system was pilot tested in 2011 at the San Francisco Medical Center of Kaiser Permanente, the largest healthcare organization in the United States. An independent evaluation was carried out at the conclusion of the test.

The present section will

- describe Converser for Healthcare and its pilot project;
- summarize the resulting evaluation;
- provide an extended example of the revised system in use; and
- discuss the principal customization facility, Translation Shortcuts.

**System Description** We begin with a brief description of Converser's approach to interactive automatic interpretation, focusing upon the system's verification, correction, and customization features.

[3] For Project DIPLOMAT, see Frederking et. al. (2000); for BABYLON, see Waibel et al. (2003); for TRANSTAC, see Frandsen et al. (2008); and for GALE, see Cohen (2007) and Olive et al. (2011). Concerning Project BOLT, see "Broad Operational Language Translation (BOLT) (Archived)." Defense Advanced Research Projects Agency (DARPA). URL: www.darpa.mil/program/broad-operational-language-translation.

First, users could monitor and correct the speech-recognition system to ensure that the text which would be passed to the MT component was completely correct. Speech, typing, or handwriting could be used to repair speech-recognition errors.

Next, during the MT stage, users could monitor – and, if necessary, correct – one especially important aspect of the translation, lexical disambiguation.

The system's approach to lexical disambiguation was twofold: First, Converser supplied a back-translation, or retranslation of the translation from the target language back to the source. Using this paraphrase of the initial input, even a monolingual user could make an initial judgment concerning the quality of the preliminary MT output. Other systems, such as IBM's MASTOR (Gao et al., 2006), have also employed retranslation. Converser, however, exploited proprietary technologies to ensure that the lexical senses used during back-translation accurately reflected those used in forward translation.

In addition, if uncertainty remained about the correctness of a given word sense, the system supplied a proprietary set of Meaning Cues™ – synonyms, definitions, and so on – which had been drawn from various resources, collated in a database (called SELECT™), and aligned with the respective lexica of the relevant MT systems. With these cues as guides, the user could monitor the current, proposed meaning and, when necessary, select a different, preferred meaning from among those available. Automatic updates of translation and back-translation then followed.

The initial purpose of these techniques was to increase reliability during real-time speech-translation sessions. Equally significant, however, they could also enable even monolingual users to supply feedback for offline machine learning to improve the system. This feedback capability remains rare: Usually, only users with some knowledge of the output language can supply it, for example, in Google's Translate Community.

All translations were recorded in bilingual transcripts, including both the original source language and the target language translation. (In the latest system versions, transcripts also contained relevant back-translations.)

Converser adopted rather than created its speech and translation components, adding value through the interactive interface elements to be explained. Nuance, Inc., later acquired by Microsoft, supplied speech recognition; rule-based English and Spanish bi-directional MT were supplied by Word Magic of Costa Rica;[4] and TTS was again provided by Nuance.

Identical facilities were available for Spanish as for English speakers: When the Spanish flag was clicked, all interface elements – buttons and menus,

[4] "Word Magic." URL: https://word-magic-translator-home-edition.software.informer.com/.

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

onscreen messages, Translation Shortcuts (Section 2.3.2.8), handwriting recognition, and so on – changed to Spanish.

**Multimodal Input** In healthcare settings, speech input isn't appropriate for every situation. Current speech-recognition systems remain unfamiliar for many users. To maximize familiarity, Converser incorporated standard commercial-grade dictation systems for broad-coverage and ergonomic speech recognition, products with established user bases in the healthcare community. Even so, some orientation and practice were required. Also expected were problems of ambient noise (e.g., in emergency rooms or ambulances) and problems of microphone and computer arrangement (e.g., to accommodate not only desktops but counters or service windows, which may form barriers between staff and patient).

To deal with these and other usability issues, Converser provided a range of input modes: Also enabled, in addition to dictated speech, were the use of touchscreen keyboards for text input and the use of standard keyboards. All of these input modes had to be bilingual, and language switching needed to be arranged automatically when there was a change of active participant. Further, it was possible to change input modes seamlessly within a given utterance: For example, users could dictate the input if they wished but then have the option to make corrections using handwriting or one of the remaining two modes.

Of course, even this flexible range of input options hardly solved all problems. Illiterate patients pose special difficulties. The careful and relatively concise style of speech required for automatic recognition is often difficult to elicit, so that recognition accuracy remains low, and the ability to read and correct the results is obviously absent. Just as obviously, the remaining three text input modes would be equally ineffectual for illiterates. Converser's approach to low literacy was to supply Translation Shortcuts for the minimally literate. It was hoped that future versions would augment Shortcuts with TTS and iconic pictures.

Staff members are usually at least minimally literate, but they present their own usability issues. Their typing skills may be low or absent. Handling the computer and microphone may be awkward in many situations, for example, when examining a patient or taking notes. (Speech-translation systems are expected to function in a wide range of physical settings: in admissions or financial aid offices, at massage tables for physical therapy with patients lying face down, in personal living rooms for home therapy or interviews, and in many other locations.)

To help deal with the awkwardness issues, one version of the system provided voice commands, enabling hands-free operation. Both full interactive

translation and the Translation Shortcuts facility could then be run hands free. To a limited degree, the system could be used eyes free as well: TTS could be used to pronounce the back-translation so that preliminary judgments of translation quality could be made without looking at the computer screen. These facilities, however, remained insufficiently tested in the pilot project to be discussed now.

**Pilot Project** In 2011, Converser for Healthcare 3.0 was pilot tested at the Medical Center of Kaiser Permanente in San Francisco. The project, supported by a grant from the company's Innovation Fund, ran for nine calendar months, with use in three departments during three of those months. At the conclusion, sixty-one interviews were conducted by an interpreter from an outside agency. A formal internal report gave the results. Reception was generally positive (Section 2.3.2.1.4); but departmental responsibility for next steps remained divided on project completion, and there has been no further use to date.

Converser was used and evaluated in four use cases in the Medical Center's Pharmacy, and one each in Inpatient Nursing and Eye Care. Each use case had its own workflow and equipment setup. In the Pharmacy, the master computer could be stationary (in the Consulting or Drop-off use case); handheld (in the Pickup use case); or on a cart (in the Greeter use case). In Inpatient Nursing, a handheld tablet personal computer was used throughout. In Eye Care, to facilitate typing, stationary use of the tablet was preferred. The hardware and software used in the project are described and assessed in Seligman and Dillinger (2011). The project's logistical issues are also discussed in detail.

**Evaluation** Evaluation of the Kaiser Permanente project relies on Kaiser's internal report, based as mentioned on a commissioned survey by an independent third party. The report itself is proprietary, but its findings are reproduced in essence in Seligman and Dillinger (2015; 2016). One significant finding: when asked whether the system met their needs, of the 79 percent of interviewed patients who answered the question, 94 percent responded either "completely" or "mostly." However, as would be expected in a system fielded a decade ago, qualifications and stumbling blocks were not lacking. The cited papers report on these, and on revisions subsequently undertaken to resolve them.

**Revised System in Use** Following is an extended example of the revised system in use, with emphasis on features addressing reliability and customization issues. For ease of exposition, we use the present tense, though the system isn't currently in use.

Again, depending on the platform, the system can offer up to four input modes: speech, typing, handwriting, and touchscreen. To illustrate the use of interactive correction for speech recognition as well as MT, we assume that the user has clicked on the round red Mic button to activate the microphone (Figure 2.1).

Still in Figure 2.1, notice the Traffic Light Icon™ and two Earring Icons™. These are used to switch between Precheck Mode and NoPrecheck Mode for translation and speech recognition, respectively. Both icons are currently green, indicating "Full speed ahead!" That is, verification has been temporarily switched off: The user has indicated that it is unnecessary to precheck either ASR or MT before transmitting the next utterance, preferring speed to accuracy.

Just prior to the figure's snapshot, the user said, "San Jose is a pleasant city." Since verification had been switched off for both ASR and MT, these functioned without interruption. The speech-recognition result appeared briefly (and in this case correctly) in the Input window. Immediately thereafter, the Spanish translation result (also correct in this case) appeared in the right-hand section of the Transcript window and was immediately pronounced via TTS. Meanwhile, the original English input was recorded in the left-hand section of that window.

Also on the English side of the Transcript window and just below the original English input is a specially prepared back-translation. The original input was translated into Spanish and then retranslated back into English. Proprietary

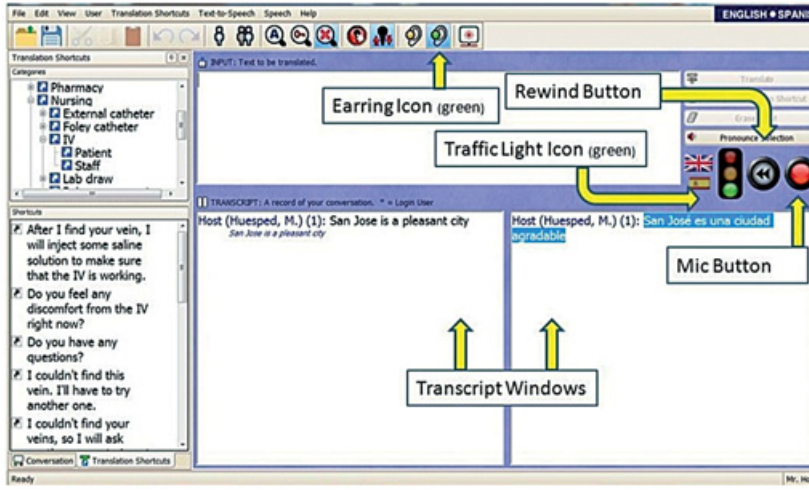


Figure 2.1 Earring and Traffic Light Icons are green: "Full speed ahead!"

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

techniques ensure that the Spanish-to-English back-translation means the same as the Spanish. Thus, even though pre-verification was bypassed for this utterance in the interest of speed, post-verification via the Transcript window was still enabled. (This window, containing inputs from both English and Spanish sides and the associated back-translations, can be saved for record-keeping. Participant identities can optionally be masked for confidentiality.)

Using this back-translation, the user might conclude that the translation just transmitted was inadequate. In that case, or if the user simply wants to rephrase this or some previous utterance, he or she can click the Rewind Button (round, with chevrons). A menu of previous inputs then appears (not shown). Once a previous input is selected, it will be brought back into the Input window, where it can be modified using any available input mode – voice, typing, or handwriting. In our example sentence, for instance, "pleasant" could be changed to "boring"; clicking the Translate button would then trigger translation of the modified input, accompanied by a new back-translation.

In Figure 2.2, the user has selected the yellow Earring Icon, specifying that the speech recognition should "proceed with caution." As a result, spoken input remains in the Input window until the user explicitly orders translation. Thus, there's an opportunity to make any necessary or desired corrections of the ASR results. In this case, the user has said "This morning, I received an email from my colleague Igor Boguslavsky." The name, however, has been misrecognized as "Igor bogus Lovsky." Typed or handwritten correction can fix the mistake, and the Translate button can then be clicked to proceed.

Just prior to Figure 2.3, the Traffic Light Icon was also switched to yellow, indicating that translation (as opposed to speech recognition) should also "proceed with caution": It should be prechecked before transmission and pronunciation. This time the user said, "This is a cool program." Since the

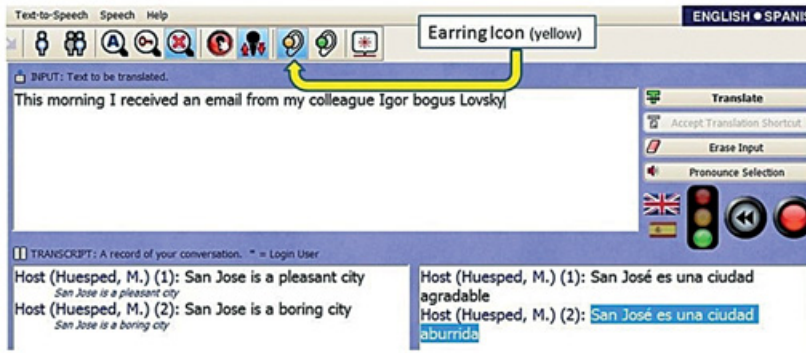


Figure 2.2 Earring Icon is yellow: "Proceed with caution!"

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

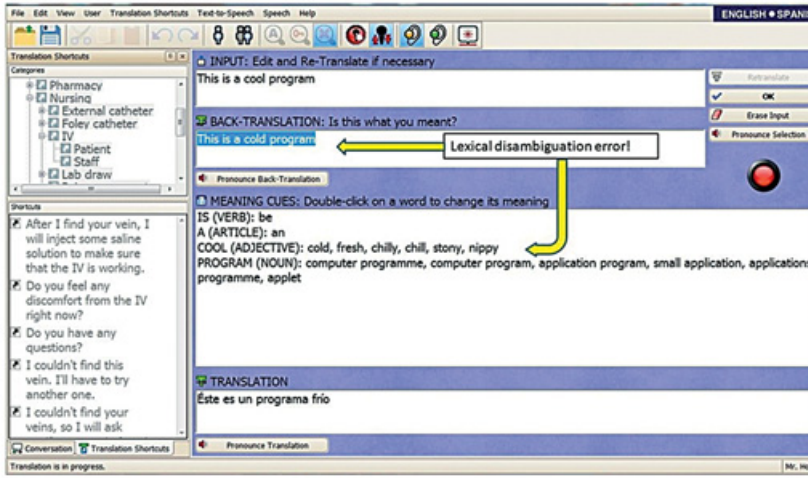


Figure 2.3 Verification Panel, with a lexical disambiguation error in "This is a cool program."

Earring Icon is still yellow, ASR results were prechecked and approved. Then the Verification Panel™ appeared, as shown in the figure. At the bottom, we see the preliminary Spanish translation, "Éste es un programa frío."

Unfortunately, despite the best efforts of the translation program to determine the intended meaning in context, "cool" has been mistranslated – as shown by the back-translation, "This is a cold program." Another indication of the error appears in the Meaning Cues window (third from the top), which indicates the meaning of each input word or expression as currently understood by the MT engine. Converser 4.0 employs synonyms as Meaning Cues; but pictures, definitions, and examples might also be used. In the present case, we see that the word "cool" has been wrongly translated as "cold, fresh, chilly . . ."

To rectify the problem, the user double clicks on the offending word or expression. The Change Meaning Window™ then appears (Figure 2.4), with a list of all available meanings for the relevant expression. Here, the third meaning for "cool" is "great, fun, tremendous . . ." When this meaning has been selected, the entire input is retranslated. This time the Spanish translation will be "Es un programa estupendo," and the translation back into English is "Is an awesome program." The user may accept this rendering, despite the minor grammatical error, or may decide to try again.

A side note concerning the Traffic Light Icon and Earring Icons: These help to balance a conversation's reliability with its speed. And again, while reliability is indispensable for serious applications like healthcare, some time is

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

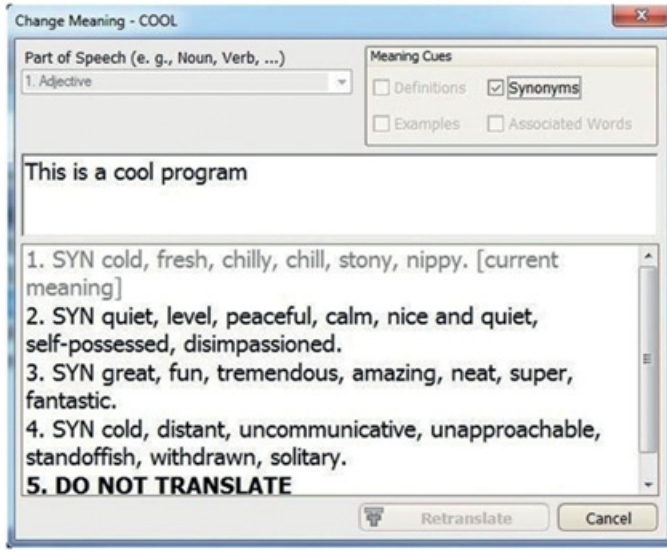


Figure 2.4 The Change Meaning Window, with four meanings of "cool"

required to interactively enhance it. The icons let users proceed carefully when accuracy is paramount, or a misunderstanding must be resolved. On the other hand, they can move ahead more quickly when throughput is judged more important. This flexibility, we anticipate, will be useful in future applications featuring automatic detection of start-of-speech: In NoPreCheck Mode, ASR and translation will proceed automatically without start or end signals, and thus without demanding the user's attention, but can be interrupted for interactive verification or correction as appropriate. (On the attention required for optional monitoring, compare Section 2.2.1.3.1.)

**Translation Shortcuts** We now shift focus from Converser's reliability features to its principal facility for customization and adaptation to multiple use cases: Translation Shortcuts – pre-packaged translations, providing a kind of translation memory. Shortcuts are designed to provide two main advantages.

First, translations have been professionally verified, so their reverification is unnecessary. They can be reliably transmitted as is. As such, they do double duty for reliability and customization.

Second, access to stored Shortcuts is very quick, with little or no need for text entry – a plus especially for busy use cases like healthcare. Several facilities contribute to meeting this design criterion.

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

- A Translation Shortcuts Browser™ is provided (on the left in Figures 2.1, 2.3, and 2.5) so that users can find needed Shortcuts by traversing a tree of Shortcut categories. Using this interface, users can execute Shortcuts, even if their ability to input text is quite limited, by tapping or clicking. Points to notice:
  - The Translation Shortcuts panel can be slid in and out of view to conserve screen space and avoid distraction. (In one Converser version, it could be operated by voice commands.)
  - The Shortcuts Browser contains two main areas, Shortcuts Categories (above) and Shortcuts List (below).
  - In the Categories section of Figures 2.1 and 2.3, the Nursing category has been selected. It contains several subcategories including External catheter, Foley catheter, IV (intravenous), and Lab draw. The IV subcategory has been expanded to show its Patient and Staff sub-subcategories, and the latter, containing expressions most likely to be used by healthcare staff members, has been selected. There is also a Patients subcategory, used for patient responses.
  - Below the Categories section is the Shortcuts List section, containing a scrollable list of alphabetized Shortcuts. (Various other sorting criteria could be enabled, for example, sorting by frequency of use, recency, etc.)
  - Double-clicking on any visible Shortcut in the list will execute it. (Clicking once will select and highlight a Shortcut, and typing Enter will execute any currently highlighted Shortcut.)
  - If a Shortcut from a Staff subcategory has been used, the associated Patient subcategory can be opened automatically to enable a response.
- A Shortcut Search™ facility can retrieve a set of relevant Shortcuts given only keywords or the first few characters or words of a string. The desired Shortcut can then be executed with a single gesture (mouse click or stylus tap) or voice command.
  - In Figure 2.5, the Mental Health category has been selected, and an icon (showing a magnifying glass containing a key) has been clicked to authorize Keyword Search.
  - The word "you" has been entered in the Input buffer – by voice, typing, or handwriting – and several Shortcuts containing this word have been found and gathered in a scrollable menu, ready for clicking.
  - Here, the results are sorted alphabetically. Various additional sorting possibilities might also be useful: by frequency of use, proportion of matched words, and so on.
  - Arrow keys or voice commands can be used to navigate the results.

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

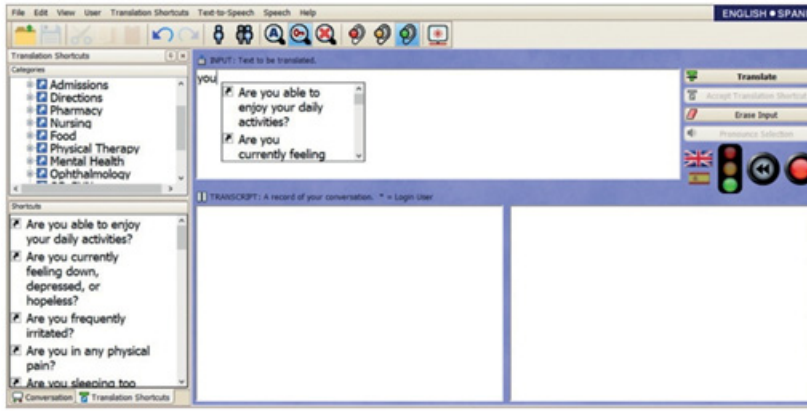


Figure 2.5 Automatic keyword search for Translation Shortcuts

- If the user enters the exact text of any Shortcut, a message will identify it as such, indicating that verification will not be necessary.
- However, final text not matching a Shortcut will be passed to the routines for full translation with verification. In this way, a seamless transition is provided between the Shortcuts facility and full, broad-coverage translation.

Again, because the Shortcuts Browser can be used without text entry, simply by pointing and clicking, it enables responses by minimally literate users. Use by completely illiterate users could be enabled through automatic pronunciation of Shortcuts and categories in the Shortcuts Browser via TTS, in effect reading the Shortcuts aloud while highlighting them. Shortcuts could also be augmented with pictures or symbols as clues to their meaning.

Having scrutinized Converser for Healthcare in terms of both reliability and customizability, we now turn to another healthcare-oriented SLT system supporting full translation. In this system, ergonomics – customization facilitating practical use in the specific settings – has been the central focus of research and development.

#### 2.3.2.2 Fujitsu's Focus on Ergonomics

At Fujitsu Laboratories Ltd., the Artificial Intelligence Laboratory recently developed a system supporting full speech translation for healthcare under the direction of Senior Researcher Tomoki Nagase. The work was carried out in cooperation with Japan's Global Communication Plan Project associated with the planned-but-canceled Tokyo Olympics in 2020. This research and development, tightly focused on practical use in the healthcare setting, exemplifies purposeful customization for the assigned application.

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

Japan was expecting some 40 million visitors to the games, and the number of foreign residents in Japan had been increasing as well. The COVID-19 pandemic disrupted both expectations, but needs for healthcare translation can be expected to resurge. Prior to the disruption, in response to Fujitsu's questionnaires, about 70 percent of the healthcare institutional respondents anticipated language problems, so the need for communication aids appeared clear, particularly for minor languages and over holidays or at night. Human interpreters would have been preferred, but service might have been inefficient due to intermittent use, so interest was strong in technology-based solutions, with due recognition of their limits.

In preparation for clinical trials, Fujitsu organized cooperation with medical and research institutions, using interviews and translation logs to gather feedback concerning design. Principal partners were the International Medical Center of the University of Tokyo Hospital (which provided ethical review no. 10704) and the National Institute of Information and Communications Technology (NICT), which supplied crucial software and pursued performance improvements through analysis of speech-translation logs. Fujitsu's responsibility was to develop terminals and interfaces to be used at medical sites.

Preliminary simulation tests clarified several points. First, hands-free solutions would be needed, to leave both hands free for medical work and to help prevent infection. Second, a fallback would be needed in case of misunderstandings unresolved by repetition. To address the first requirement, two solutions were developed: (1) a fixed desktop terminal, with which staff and patient could interact face-to-face over a desk or counter, as in reception areas, medicine or cashier counters, blood sampling or inspection stations, and so on, and (2) a wearable terminal, usable by staff responding to foreign nationals in hospital wards, nursing stations, and so on. Tests confirmed stable operation in various noisy environments. To meet the fallback need, both terminals were equipped with a button for calling up a human interpreter.

Following these preparations, clinical trials were undertaken in 2016 and 2017, starting with six and progressing to twenty-one hospitals. English and Japanese were handled via the desktop terminal throughout, with Chinese>Japanese and wearable terminals added in the second year of trials. Use cases were selected freely by the institutions, without restrictions on conversations: Reception, hospital wards, and examinations were the most frequent users, along with medical interviews, intensive care units, inspections, medicine counters, emergency visits, cashiers, examination or treatment sessions, and others. Consent signatures were obtained from patients, and sessions were followed up with optional questionnaires for staff members and patients.

During the clinical trials, eighty-three English–Japanese sessions and seventy-six Chinese>Japanese sessions were recorded. Perhaps not surprisingly, speech

was initiated by medical staff twice as often as by patients (67 percent compared to 33 percent). And, interestingly, there were more Chinese- than English-speaking users (53 percent and 47 percent). The optional questionnaire posed four questions to staff and patients: Was it useful during the conversation? Was what you spoke understood? Did you understand what the other person spoke? Was it easy to use? Five degrees of satisfaction could be registered, from "Highly rated" to "Lowly rated."

Combined scores for "Highly" or "Reasonably" ranged over these questions from about 60 percent to 70 percent for staff members and from about 70 percent to 75 percent for patients. The best result: About 60 percent of the English speakers responded that their understanding of Japanese was "Highly rated." By comparison, the "Highly rated" score for Chinese patients' comprehension of Japanese was about 35 percent.

With respect to the wearable terminal, in a briefing before clinical trials began, fifty staff members were interviewed. Forty-five, or 90 percent, said they were able to converse effectively. The terminal's size and weight were generally judged acceptable. Asked if they'd want to use the terminal at work, twenty-three said yes, as soon as possible; sixteen said they'd wait until the translation accuracy was improved; and the remainder would wait for an improved terminal or preferred not to use the system.

Actual trials followed improvements in the terminals, based on lessons learned. A number of positive staff reactions were claimed. Users said they were able to convey technical terms more easily with the device than with gestures; that they had more opportunities to converse with foreign patients and felt less hesitant to speak with them; and that they felt a sense of safety because the systems were available, even if there were few actual opportunities to use them.

Commercialization and deployment of the system remained for the future, but, until the disruption caused by the pandemic, plans were under way for expansion of language coverage, for example, to Korean, Vietnamese, and Brazilian Portuguese. Also anticipated were improvements in translation accuracy, especially for Chinese. Tools for training staff users, perhaps including instructional videos, were to be considered as well.

## 2.4 Past and Current Speech Translation Systems

We next point toward a range of further speech translation solutions available now or in the past, each supporting a subset of the reliability and customization features we've considered. Several useful studies and surveys will be cited.

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

### 2.4.1 Reliability of Machine Translation for Healthcare: A Study

As a component of speech-to-speech translation systems, translation technology development has been especially dynamic, with several changes of basic approach (Chapter 1). The consequent improvement in raw translation *accuracy* has brought improvement in speech translation *reliability* (accuracy plus user confidence), a critical factor in widespread adoption. But how much improvement?

Sample translations and back-translations were appended to Chapter 1 to give an informal impression of the state of the art in automatic text translation. We can now mention a pertinent formal study of the translation system sampled there: an evaluation of its Japanese-to-English translation in the medical domain (Takakusagi et al., 2021). Interestingly, back-translation into the original Japanese also figured prominently in this research.

The system in question is DeepL Translator, developed by DeepL GmbH, Cologne, Germany.[5] The test case was an already-published medical article in Japanese, automatically translated into English using DeepL Translator. The resulting English article was then back-translated into Japanese by three researchers. Three other researchers then compared the back-translated Japanese sentences with the original Japanese manuscript and calculated the percentage of sentences keeping the intended meaning. The match rate for the article as a whole was found to be 94.0 ± 2.9 percent. Different sections of the article fared differently, with significantly higher rate in the Results section, but lower rates in the Methods section. Helpfully, however, significant predictors for mismatched translations were found, with the most mismatches in compound sentences and sentences with unclear subjects and predicates. (Chapter 3 studies the usefulness of such predictors.) Overall, the translation was judged accurate.

While the system apparently delivered translation results in the 90 percent-plus range for *written* material – *and* on a famously challenging translation direction, *and* with a translation system not specifically trained on medical material – the added difficulties of translating recognized text from spontaneous speech must be considered (Chapter 1). Even so, from the translation viewpoint, the prospects for future speech translation systems do seem quite promising, especially since the usability of back-translation for verification and optional correction has already been demonstrated for at least this particular translation component. And so, with cautious optimism, we go on to refer readers to several surveys of speech translation systems.

[5] "DeepL Translator." *Wikipedia*, Wikimedia Foundation, 10 August 2022, at 17: 37(UTC), https://en.wikipedia.org/wiki/DeepL_Translator.

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

### 2.4.2 Surveys of Speech Translation Systems

"Enabling Medical Translation for Low-Resource Languages" (Musleh et al., 2016) briefly describes some speech translation systems available at the time of writing while developing text translation for Urdu, an under-resourced language closely related to Hindi and important for healthcare in Qatar. The paper provides useful historical context, even as several of the surveyed systems remain active.

#### 2.4.2.1 Some Bi-directional Speech Translation Systems

The first group of systems cited by Musleh and colleagues are those for bi-directional doctor-patient communication, with special interest in systems requiring data collection for under-resourced languages. Most built until the time of writing (Bouillon et al., 2008; Dillinger and Seligman, 2006; Eck et al., 2010; Ehsani et al., 2006; Gao et al., 2006; Heinze et al., 2006) remained prototypes, with few fully deployed. Some did, however, work with under-resourced languages (Bouillon et al., 2008; Ehsani et al., 2006; Heinze et al., 2006; Gao et al., 2006). These relied on symbolic meaning representations rather than on statistical machine translation (while neural translation remained in in the future). Unfortunately, none of the systems addressed the top five languages of most interest to Qatar. In addition to Converser for Healthcare and S-MINDS, already discussed, the following systems are cited:

- MedSLT (Bouillon et al., 2008), an interlingua-based speech-to-speech translation system, covering a restricted set of domains for English, French, Japanese, Spanish, Catalan, and Arabic. The doctors' questions or statements to the patient could be translated, but not the patients' responses.
- Jibbigo (Eck et al., 2010), a travel and medical speech-to-speech MT system, deployed on iPhone mobile application (and requiring no Internet connection). Jibbigo covered English<>Spanish for medical translation.
- Accultran (Heinze et al., 2006), a prototype featuring back-translation to the doctor for confirmation and yes/no or multiple-choice questions to the patient. A cross-cultural adviser was included. Sensitive and hard-to-translate utterances were flagged. The SNOMED-CT or Clinical Document Architecture (CDA-2) standards were used as an interlingua.
- IBM MASTOR (Gao et al., 2006), a speech-to-speech MT system for English<>Mandarin and English<>Arabic dialects. Laptops and handhelds were accommodated.
- English-Portugese SLT (Santos Gomez Rodrigues, 2013), an English-Portuguese speech-to-speech system, usable as an online service or as a mobile app.

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

#### 2.4.2.2 Some Phrase-Based Speech Translation Systems

The second group of systems discussed by Musleh et al. included several phrase-based mobile or web applications for doctor-to-patient translation only. The most popular were UniversalDoctor, MediBabble, Canopy, MedSpeak, MavroEmergency Medical Spanish, and DuoChart.[6] None enabled full (free, unseen, or spontaneous) translations, and none covered the language pairs of interest for Qatar. Some (e.g., UniversalDoctor) required paid subscriptions.

#### 2.4.2.3 Fifteen Representative Apps: A Study

"Language Translation Apps in Health Care Settings: Expert Opinion" (Panayiotou et al., 2019) offers an assessment of fifteen apps. The concentration was on iPad-compatible language translation apps: Were they suitable for everyday conversations in healthcare settings? Apps found on the Apple iTunes Store and in the literature were considered if available free and able to translate at least one of the top ten languages spoken in Australia. These were reviewed in two stages: Stage 1 entailed a feature analysis by two independent researchers, with evaluation for offline use, input and output methods, and available languages; in Stage 2, two independent professionals with expertise in translation and cross-cultural communication analyzed app suitability for everyday communication in healthcare. Importantly, however, apps were considered unsuitable if they aimed at aspects of care for which professional interpreters were normally responsible. These included assessment, treatment and discharge planning, and elicitation of consent for medical treatments.

Eight of the fifteen evaluated apps contained voice-to-voice and voice-to-text translation options. Six were phrase-only systems, and one supplied a combination of free input and preset phrases. Five apps were excluded before Stage 2. Of the ten remaining apps, six were specifically designed for healthcare translation purposes. Of these, two were rated as suitable for everyday communication in the healthcare setting: Assist and Talk to Me. Both were found to be culturally and linguistically diverse and to contain simple and appropriate preset health phrases. Neither attempted conversations normally handled by professional interpreters.

The study concluded cautiously: All iPad-compatible translation apps require caution and consideration in healthcare settings, and none should

[6] "Universal Doctor." URL: www.universaldoctor.com; "Medibabble." URL: http://medibabble.com; "Canopy." URL: www.canopyapps.com; "Medspeak." URL: https://apptopia.com/ios/app/313250795/about; "MavroEmergency Medical Spanish." URL: http://mavroinc.com/medical.html; "DuoChart." URL: http://duochart.com.

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

replace professional interpreters. However, a few apps were found suitable for everyday conversations, especially phrase-based systems treating subjects not requiring a professional interpreter.

#### 2.4.2.4 Some Additional Links

Finally, several additional healthcare-related studies have been kindly suggested by Meng Ji, co-author of this volume: Van de Velde et al. (2015); Thonon et al. (2021); Chen et al. (2017); and Turner et al. (2019).

## 2.5 Conclusions

This chapter's introduction promised an optimistic conclusion concerning the future of speech translation systems for healthcare. Optimism is certainly warranted – firstly, in view of recent technological progress, not only in the speech and translation components but in the related ecosystem of platforms, peripherals, and security; and secondly, considering the prospects for continued improvements in the associated reliability and customizability.

However, as (almost) goes without saying, technological optimism must be tempered by prudent and informed caution – especially in healthcare, where errors can be deadly. While we do advocate progressive adoption of speech translation technologies in many healthcare-related use cases, we do so with these caveats:

- Staff responsible for tech selection require basic grounding in the relevant tech: How does it work, and what are its limitations per use case? We hope that this volume can help to supply that foundation. While the technology is challenging and quickly developing, it should not be treated as oracular. While systems unavoidably remain black boxes to some extent in current stages of the neural network era, blind or awestruck acceptance is unhealthy – in healthcare, quite literally.
- Further, while responsible staff should strive for at least high-level understanding, they shouldn't fly solo. Professionals in the prospective technologies must also be consulted, with reference to specific intended use cases.
- All responsible parties – staff, consultants, and patients – must be helped to fully understand that speech recognition and translation errors are inevitable in automatic systems of whatever quality, since even human interpreters make mistakes. An aforementioned tradeoff must be acknowledged: the broader the coverage of the system, the less its expected accuracy. And so, if the use case is inherently narrow (as, e.g., for patient intake) and demands

https://doi.org/10.1017/9781108938976.003 Published online by Cambridge University Press

reliability with little staff interaction or monitoring, phrase-based rather than full translation systems may be sensible choices. As continued improvement in accuracy raises the reliability of full translation systems, or as increased staff interaction becomes possible or desirable, systems providing broader translation can be reconsidered.

Arthur C. Clarke said, "Any sufficiently advanced technology is indistinguishable from magic" – and he was not far wrong. But as we've seen, for healthcare and many other challenging use cases, "Magic is not enough" – not without determined attention to reliability and customization. Still, we find ourselves in the unaccountably fortunate position of not only witnessing but – to some degree, anyway – actually understanding developments that would have seemed purely magical even in recent decades. So while blind or awe-struck adoption of speech translation services is not recommended, awed appreciation, with eyes wide open, definitely is. We sorcerers' apprentices would be ungrateful not to exploit this sorcery to improve well-being and save lives. But with care.

## References

Bouillon, Pierrette, Glenn Flores, Maria Georgescul, et al. 2008. "Many-to-Many Multilingual Medical Speech Translation on a PDA." In *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas*, AMTA'08, pages 314–323, Waikiki, Hawaii, USA.

Chen X, S. Acosta, and A. Barry. 2017, "Machine or Human? Evaluating the Quality of a Language Translation Mobile App for Diabetes Education Material." *Journal of Medical Internet Research*, 2(1): e13. URL: https://diabetes.jmir.org/2017/1/e13. DOI: 0.2196/diabetes.7446.

Cohen, J. 2007. "The GALE Project: A Description and an Update." In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding* (ASRU), Kyoto, Japan.

Dillinger, M., and M. Seligman. 2004. "System Description: A Highly Interactive Speech-to-Speech Translation System." In *Proceedings of the Association for Machine Translation in the Americas* (AMTA-04). Washington, DC.

Dillinger, M., and M. Seligman. 2006. "Converser: Highly Interactive Speech-to-Speech Translation for Healthcare." In *Proceedings of the COLING-ACL 2006 Workshop on Medical Speech Translation*, pages 36–39, Sydney, Australia.

Eck, M., I. Lane, Y. Zhang, and A. Waibel. 2010. "Jibbigo: Speech-to-Speech Translation on Mobile Devices." In *Proceedings of IEEE Spoken Language Technology Workshop*, SLT2010, pages 165–166, Berkeley, California, USA.

Ehsani, F., J. Kimzey, D. Master, et al. 2006. "Speech to Speech Translation for Medical Triage in Korean." In *Proceedings of the COLING-ACL 2006 Workshop on Medical Speech Translation*, pages 13–19, New York, NY, USA.

Ehsani, F., J. Kimzey, E. Zuber, D. Master, and K. Sudre. 2008. "Speech to Speech Translation for Nurse Patient Interaction." In *Coling 2008: Proceedings of the Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications*, pages 54–59, Manchester, England: COLING 2008 Organizing Committee.

Frandsen, M. W., S. Z. Riehemann, and K. Precoda. 2008. "IraqComm and FlexTrans: A Speech Translation System and Flexible Framework." In *Innovations and Advances in Computer Sciences and Engineering*, pages 527–532, Singapore: Springer.

Frederking, R., A. Rudnicky, C. Hogan, and K. Lenzo. 2000. "Interactive Speech Translation in the DIPLOMAT Project." *Machine Translation*, 15, pages 27–42.

Gao, Y., Liang G., B. Zhou, et al. 2006. "IBM MASTOR System: Multilingual Automatic Speech-to-Speech Translator." In *Proceedings of the COLING-ACL 2006 Workshop on Medical Speech Translation*, pages 53–56, Sydney, Australia.

Heinze, D. T., A. Turchin, and V. Jagannathan. 2006. "Automated Interpretation of Clinical Encounters with Cultural Cues and Electronic Health Record Generation." In *Proceedings of the COLING-ACL 2006 Workshop on Medical Speech Translation*, pages 20–27, Sydney, Australia.

Marking, M. 2021. "Zoom Bolts on Speech Translation in What Is Only Its Second-Ever Acquisition." *URL*: https://slator.com/zoom-bolts-on-speech-translation-in-what-is-only-its-second-ever-acquisition/. June 30, 2021.

Musleh, A., N. Durrani, I. Temnikova, S. Vogel, and O. Alsaad. "Enabling Medical Translation for Low-Resource Languages." 2016. In *Proceedings of 17th International Conference on Intelligent Text Processing and Computational Linguistics*, Konya, Turkey.

Mutal, J., P. Bouillon, J. Gmerlach, P. Estrella, and H. Spechbach. 2019. "Monolingual Backtranslation in a Medical Speech Translation System for Diagnostic Interviews – A NMT Approach." In *Proceedings of the MT Summit XVII: The 17th Machine Translation Summit*, Dublin, Ireland.

Olive, J., C. Christianson, and J. McCary (eds.). 2011. *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. New York, NY: Springer Science and Business Media.

Panayiotou, A., A. Gardner, S. Williams, et al. 2019. "Language Translation Apps in Health Care Settings: Expert Opinion." *JMIR Mhealth Uhealth*, 7(4), April 9, 2019. e11316. DOI: 10.2196/11316. PMID: 30964446; PMCID: PMC6477569.

Paras, M., O. Leyva, T. Berthold, and R. Otake. 2002. "*Videoconferencing Medical Interpretation: The Results of Clinical Trials*." Oakland, CA: Health Access Foundation.

Santos Gomes Rodrigues, J. A. 2013. *Speech-to-Speech Translation to Support Medical Interviews*. PhD thesis, Universidade de Lisboa, Portugal.

Seligman, M. 2020. "Socially Significant Applications of Speech-Translation Technology." In *The Oxford Handbook of Translation and Social Practices*. M. Ji and S. Laviosa, eds. Oxford University Press (Oxford Handbooks). New York. December 15, 2020. Chapter 27, pages 561–586.

Seligman, M., and M. Dillinger. 2006a. "Usability Issues in an Interactive Speech-To-Speech Translation System for Healthcare." In *HLT/NAACL-06: Proceedings of the Workshop on Medical Speech Translation*, pages 1–4, Stroudsburg, PA: Association for Computational Linguistics.

Seligman, M., and M. Dillinger. 2006b. "Converser: Highly Interactive Speech-to-Speech Translation for Healthcare." In *HLT/NAACL-06: Proceedings of The Workshop on Medical Speech Translation*, pages 36–39, Stroudsburg, PA: Association for Computational Linguistics.

Seligman, M., and M. Dillinger. 2008. "Rapid Portability Among Domains in an Interactive Spoken Language Translation System." In *Coling 2008: Proceedings of the Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications*, pages 40–47, Manchester, England: Coling 2008 Organizing Committee.

Seligman, M., and M. Dillinger. 2011. "Real-time Multi-Media Translation for Healthcare: A Usability Study." In *Proceedings of the 13th Machine Translation Summit*, Xiamen, China.

Seligman, M., and M. Dillinger. 2012. "Spoken Language Translation: Three Business Opportunities." In *Proceedings of the Association for Machine Translation in the Americas* (AMTA-12), San Diego, CA.

Seligman, M., and M. Dillinger. 2013. "Automatic Speech Translation for Healthcare: Some Internet and Interface Aspects." In *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence* (TIA-13), Paris, France.

Seligman, M., and M. Dillinger. 2014. "Behind the Scenes in an Interactive Speech Translation System." In *Proceedings of the Association for Machine Translation in the Americas* (AMTA-14), Vancouver, BC, Canada.

Seligman, M., and M. Dillinger. 2015. "Evaluation and Revision of a Speech Translation System for Healthcare." In *Proceedings of the 12th International Workshop on Spoken Language Translation*, Da Nang, Vietnam.

Seligman, M., and M. Dillinger. 2016. "Automatic Interpretation for Healthcare." *MultiLingual Computing*, pages 38–42.

Seligman, M., and A. Waibel. 2019. "Advances in Speech-To-Speech Translation Technologies." In *Advances in Empirical Translation Studies*, M. Ji, ed., pages 217–251, Cambridge, England: Cambridge University Press.

Spechbach, H., and P. Bouillon. 2019. "BabelDr – An Innovative and Reliable Translation Tool." In *Proceedings of the 18th European Congress of Internal Medicine*, Lisbon, Portugal.

Takakusagi, Y., T. Oike, K. Shirai, et al. 2021. "Validation of the Reliability of Machine Translation for a Medical Article from Japanese to English Using DeepL Translator." *Cureus*, September 6; 13 (9): e17778. DOI: 10.7759/cureus.17778. PMCID: PMC8494522.

Thonon F, S. Perrot, A. Yergolkar, et al. 2021. "Electronic Tools to Bridge the Language Gap in Health Care for People Who Have Migrated." *Systematic Review, Journal of Medical Internet Research*, 23 (5): e25131 URL: www.jmir.org/2021/5/e25131. DOI: 10.2196/25131.

Turner A, Y. Choi, K. Dew, et al. 2019. "Evaluating the Usefulness of Translation Technologies for Emergency Response Communication: A Scenario-Based Study." *Journal of Medical Internet Research, Public Health Survey*, 5 (1): e11171. URL: https://publichealth.jmir.org/2019/1/e11171. DOI: 10.2196/11171.

Van de Velde, S., L. Macken, K. Vanneste, et al. 2015 "Technology for Large-Scale Translation of Clinical Practice Guidelines: A Pilot Study of the Performance of

a Hybrid Human and Computer-Assisted Approach." *JMIR Medical Informatics*, 3(4), e33. URL: https://medinform.jmir.org/2015/4/e33. DOI: 10.2196/medinform.4450.

Waibel, A., A. Badran, A. W. Black, et al. 2003. "Speechalator: Two-way Speech-To-Speech Translation on a Consumer PDA." In *EUROSPEECH-2003, the Eighth European Conference on Speech Communication and Technology*, pages 369–372, Baixas, France: International Speech Communication Association.

Zong, C., and M. Seligman. 2005. "Toward Practical Spoken Language Translation." *Machine Translation*, 19(2), pages 113–137.