# How well do real-time machine translation apps perform in practice? Insights from a literature review

**Mark Pluymaekers**
Zuyd University of Applied Sciences
Brusselseweg 150, 6217 HB Maastricht (NL)
mark.pluymaekers@zuyd.nl

## Abstract

Although more and more professionals are using real-time machine translation during dialogues with interlocutors who speak a different language, the performance of real-time MT apps has received only limited attention in the academic literature. This study summarizes the findings of prior studies (N = 34) reporting an evaluation of one or more real-time MT apps in a professional setting. Our findings show that real-time MT apps are often tested in realistic circumstances and that users are more frequently employed as judges of performance than professional translators. Furthermore, most studies report overall positive results with regard to performance, particularly when apps are tested in real-life situations.

## 1 Introduction

In 1997, Mark Seligman wrote that “the Internet offers a tremendous opportunity for experiments with real-time machine translation (MT) of dialogues” (Seligman, 1997). In December of the same year, SYSTRAN and AltaVista launched “the first widely available, real-time, high-speed and free translation service on the Internet” (Yang & Lange, 1998). Now, 25 years later, the Google Translate app has been downloaded more than 1 billion times from the Google App Store (Pitman, 2021). Since 2011, the app offers a conversation mode, which enables users to have utterances within a dialogue translated in real-time so that their conversation partners can understand them. Other apps such as iTranslate, TripLingo and Microsoft Translator can also be used to support synchronous dialogue between interlocutors who do not speak the same language (Tao, 2022).

To the best of our knowledge, there are no publicly available data on the frequency with which MT apps are used for real-time translation and the contexts in which this occurs. However, given the popularity of these apps, it can be expected that a large number of synchronous dialogues are translated every day, and that this happens not only in informal situations, but also in professional contexts. This raises the question of how well real-time MT apps perform in these kinds of situations. Traditionally, the academic literature has paid more attention to the quality of written translations that have been produced using MT than to the output of real-time MT apps. This study aims to boost research into the performance of real-time MT apps by summarizing the findings of earlier studies in which the performance of such apps was evaluated in a professional context.

## 2 MT quality assessment

The quality of MT output has been a hotly debated topic for decades, and a wide variety of methods for its assessment have been proposed (cf. Castilho et al., 2018). When classifying these methods, authors commonly distinguish between automated metrics and human metrics (e.g., Rivera-Trigueros, 2021; Chatzikoumi, 2020). Automated metrics include Word Error Rates (WERs), precision, recall, and BLEU scores, all of which are calculated on the basis of a comparison between MT output and a reference translation created by a professional human translator.

Human metrics are further subdivided by Chatzikoumi (2020) into metrics in which human experts express a direct judgement concerning the translation quality and metrics in which no direct judgement is expressed. When experts are asked to indicate the adequacy or fluency of a machine translated text on a 5-point scale, for example, they make an explicit quality judgement. When, on the other hand, they classify the translation errors occurring in the MT output, they provide useful information for improving the application without explicitly judging the quality of the output. Measuring the post-editing effort required to reach an acceptable quality level for the target text (e.g. Lacruz et al., 2014) also provides an indirect indication of MT quality.

There are several reasons why most of the metrics discussed above can be considered less suitable for assessing real-time MT that is used to support synchronous dialogues. First of all, post-editing does not occur in such situations, so post-editing effort cannot be used as a quality indicator. In the absence of a human-generated reference translation, automated metrics can also not be calculated. Technically speaking, human experts could judge the quality of the output after the dialogue has taken place, but they would be at a disadvantage due to the limited length and disfluent nature of the source texts, particularly when speech input is used (Przybocki et al., 2011).

Moreover, it is important to acknowledge that MT quality assessment can have different purposes. Many of the metrics above were primarily developed to identify areas of improvement for MT applications that are 'under construction' (Dorr et al., 2011). For professionals contemplating the use of real-time MT in their daily professional routines, however, improving the application is not the main priority. They want to know whether using MT will enhance the quality of their interactions with patients, students or business partners who speak a different language. In some cases, they might even wonder whether the use of MT is ethically responsible given the prevalence of errors in MT output and the potentially damaging consequences of such errors in certain contexts (Vieira et al., 2020).

Taken together, these considerations suggest that the evaluation of real-time MT might best be approached from the perspective of 'fitness for purpose', which is achieved when the quality of a translation is 'good enough' for the end user to understand the information content and pragmatic intent of a translated message (Moorkens et al., 2018; Directorate General for Translation, 2016). Although this concept has featured prominently in both practical and academic discourse about translation quality for quite some time (Jiménez-Crespo, 2018), it is not yet standard practice to ask end users to assess the quality of (post-edited) MT output (cf. Van Egdom & Pluymaekers, 2019).

This raises the question to what extent existing studies into the performance of real-time MT apps are guided by the concept of fitness for purpose, and how fitness for purpose is operationalized in evaluation methods used in these studies. For the current paper, we are specifically interested in the answers to the following questions:

RQ1: To what extent are real-time MT applications tested in authentic professional situations?

RQ2: Which quality indicators are most commonly used and how are they operationalized?

RQ3: Who judges the performance of real-time MT apps?

RQ4: Which overall picture concerning the performance of real-time MT apps emerges from the research conducted so far?

We hope to find these answers by conducting a systematic literature review of prior studies (N = 34) which report an evaluation of a real-time MT app that was or could be used to facilitate a synchronous dialogue between interlocutors who did not speak the same language. More information about our methodology is provided in the next chapter.

## 3 Method

For our literature review, we collected papers published in peer-reviewed journals or conference proceedings which assessed the quality of linguistic material that was translated in real-time by an MT application and that was related to actual or potential dialogues in professional settings (e.g., healthcare, education or tourism). Studies that focused on other types of linguistic material (e.g., websites or leaflets) or only described a real-time MT system without reporting an evaluation were excluded from the sample. Subsequently, the studies included in the sample were coded on a number of key variables derived from the research questions stated above. The following sections describe the sampling method, the coding procedure and the statistical analyses.

### 3.1 Sampling

In compiling the sample, we followed a multi-step approach (see Figure 1). First, we conducted an initial search in four scientific databases (EBSCOhost, PubMed, Web of Science and Google Scholar), which were selected for reasons of practicality (i.e., accessibility via the university library) as well as quality (cf. Creswell, 2014; Gusenbauer & Haddaway, 2020). In each database, we used the following Boolean combination of search words:

("mobile translat*" OR "real-time translat*" OR "automatic translat*") OR ("translat* tool" OR "translat* app") AND ("quality" OR "evaluation" OR "usability") NOT "knowledge translation"

Depending on the search functionalities of the database, this query was applied to the abstract, the title and the abstract, or the entire text. The relevance of the articles that came up in the search results was assessed in two steps. On the basis of the abstracts, 23 articles were marked as potentially relevant. After reading the complete articles, we decided that 10 of them indeed corresponded to the inclusion criteria outlined above.

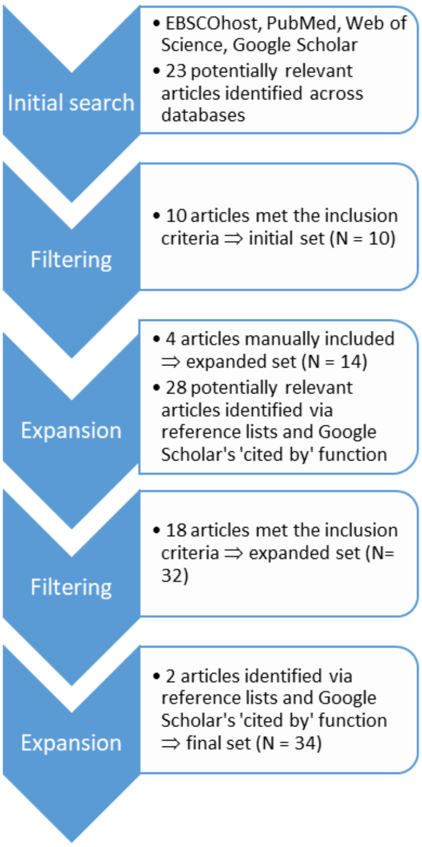
Initial search
• EBSCOhost, PubMed, Web of Science, Google Scholar
• 23 potentially relevant articles identified across databases
Filtering
• 10 articles met the inclusion criteria ⇒ initial set (N = 10)
Expansion
• 4 articles manually included ⇒ expanded set (N = 14)
• 28 potentially relevant articles identified via reference lists and Google Scholar's 'cited by' function
Filtering
• 18 articles met the inclusion criteria ⇒ expanded set (N= 32)
Expansion
• 2 articles identified via reference lists and Google Scholar's 'cited by' function ⇒ final set (N = 34)

**Figure 1:** Overview of the sampling process

In the next step, we expanded the sample by (1) manually adding 4 articles that we had found earlier and (2) investigating studies that were either included in the reference list of one of the articles in the initial set or that referred to one of the articles in the initial set. By doing so, we identified 28 potential additions to the sample, 18 of which met the screening criteria. For the newly added articles (4+18), we repeated the reference check described above, which led to the identification of 2 more articles. After this, saturation was achieved, resulting in a final sample of 34 articles (see Appendix A). More information about the characteristics of these articles (year of publication, the number and types of applications tested, language combinations etc.) will be provided in section 4.1 below.

### 3.2 Coding

All articles were coded by two independent coders using the coding scheme presented in Table 1.

| | |
|---|---|
| Year of publication | |
| Publication type | ☐ Conference paper<br>☐ Journal article |
| Professional domain | ☐ Healthcare<br>☐ ICT<br>☐ Education<br>☐ Tourism<br>☐ Other, namely: |
| # of applications | |
| Application type | ☐ Existing generic<br>☐ Existing domain-specific<br>☐ Tailor-made |
| Modality | ☐ Text-to-text<br>☐ Text-to-speech<br>☐ Speech-to-text<br>☐ Speech-to-speech |
| Language combination(s) | |
| Test type(s) | ☐ Real-life situation<br>☐ Scenario-based simulation<br>☐ Corpus-based simulation |
| Data collection method(s) | ☐ Survey<br>☐ Interview<br>☐ Focus group<br>☐ Content analysis<br>☐ Observation<br>☐ Other, namely: |
| Judge(s) | ☐ Provider<br>☐ Recipient<br>☐ User (no provider-recipient relationship)<br>☐ Professional translator<br>☐ Native speaker / bilingual<br>☐ Other, namely: |

| Quality indicator(s) | # | **Variable** | **Operationalization** |
|---|---|---|---|
| | 1 | | |
| | 2 | | |
| | 3 | | |
| Overall evaluation | ☐ Positive<br>☐ Negative<br>☐ Mixed | | |

**Table 1:** Coding scheme

Any disagreements between the two coders were discussed until consensus was reached. Most variables in the table are more or less self-explanatory, but there are three variables we wish to elaborate on here. First of all, *application type* was included to be able to distinguish between MT applications created for general purposes (e.g., Google Translate), MT applications created for specific professional domains (e.g., Canopy Medical Translator) and MT applications created by the authors of the article. With respect to *test type*, we noticed during the screening process that not all applications are tested in situations that involve actual dialogue; Sometimes, frequently occurring utterances from professional dialogues are provided to the application to assess the quality of the translation (referred to as 'corpus-based simulation' in Table 1). If actual dialogues are involved in the test, they can be either real-life dialogues or dialogues from a role-playing scenario scripted by the researchers. Finally, for the variable *judge* we decided to distinguish between providers and recipients of care, service or education, as our initial observations suggested that providers may be asked more frequently to assess the performance of MT apps than recipients.

### 3.3 Analysis

The outcomes of the coding process were entered into an SPSS data file containing mainly nominal variables recording the presence or absence of certain methodological features (e.g., whether recipients were asked to judge the performance of the app or whether focus groups were used to collect data). To gain insight into the sample characteristics and answer the research questions, frequency tables were created. To assess whether the overall judgement regarding the performance of the app differed as a function of methodological choices made, we used Chi-squared tests.

## 4 Results

### 4.1 Sample characteristics

All studies in the sample were published between 2005 and 2022. Figure 2 shows how the studies were distributed over the years. 28 studies (82%) were published in peer-reviewed journals, while 6 (18%) appeared in conference proceedings. The majority of the studies (27 or 79%) focused on one real-time MT application; 5 studies (15%) made a comparison between two applications while only 2 studies (Hwang et al., 2022 and Panayiotou et al., 2020) included three applications in their evaluation. Existing general-purpose applications were tested most frequently (18 studies or 53%), followed by apps that were created by the authors themselves and existing domain-specific applications, which were tested in 12 (35%) and 8 (24%) studies respectively. Most evaluations were conducted in the context of healthcare (28 studies or 82%). A wide variety of tested language combinations could be observed in the sample, although the majority of studies (24 or 71%) looked at one or two combinations, and English was part of the tested language combinations in 25 of the 34 studies (74%).

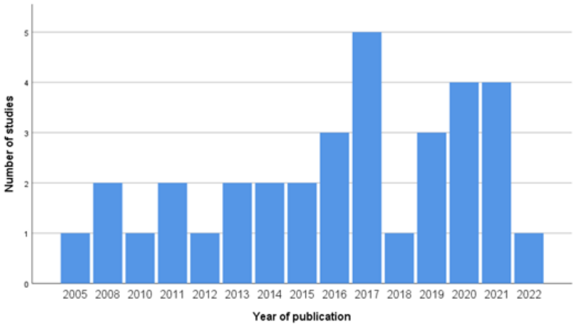

**Figure 2:** Number of studies by year of publication

### 4.2 Test types and data collection methods

Of the 34 studies in the sample, 32 used a single test type. The two exceptions were Calefato et al. (2016) and Haith-Cooper (2014), who conducted both a scenario-based and a corpus-based simulation. Far more common was the use of multiple data collection methods, which was observed in 18 of the 34 studies (53%). Tables 2 and 3 show which test types and data collection methods were used most frequently.

As Table 2 shows, most studies made an attempt to conduct a test in more or less authentic circumstances, be it in real life or during a scenario-based simulation. As can be seen in

Table 3, quantitative data collection methods such as surveys, content analysis (e.g., counting the number of correctly translated words or sentences) and observation (e.g., measuring how long it took participants to accomplish a certain task) were more popular than qualitative data collection methods, such as interviews and focus groups.

| Test type | Number of studies |
|---|---|
| Real-life situation | 16 (47%) |
| Scenario-based simulation | 15 (44%) |
| Corpus-based simulation | 5 (15%) |

**Table 2:** Test types and the number of studies they were used in (including percentages)

| Data collection method | Number of studies |
|---|---|
| Surveys | 23 (68%) |
| Content analysis | 13 (38%) |
| Observation | 12 (35%) |
| Interviews | 8 (24%) |
| Focus groups | 3 (9%) |

**Table 3:** Data collection methods and the number of studies they were used in (including percentages)

### 4.3 Quality indicators and judges

The majority of the studies (27 or 79%) employed multiple quality indicators to assess the performance of the MT app(s) under study. For judges, this was not the case, as 20 studies (59%) relied on a single category of judges. The quality indicator used most often was usability or ease of use, although it was used in only half of the studies in the sample. Similarly, providers were the most frequently employed judges, but they were still only involved in 18 out of the 34 studies (53%). Tables 4 and 5 summarize the frequency information for the different quality indicators and categories of judges.

Table 4 shows that many different quality indicators were used, some of which showed conceptual overlap even though they were referred to using different terms. That is why we decided to group them together in the table. It should be noted, however, that many studies did not provide explicit definitions of their quality indicators and that there was little uniformity in the way that variables such as *ease of use* or *accuracy* were measured. With respect to the judges, providers were more frequently asked to provide their opinion than recipients, and professional translators were involved in only a handful of studies.

| Quality indicator | Number of studies |
|---|---|
| Usability / ease of use | 17 (50%) |
| Accuracy / adequacy / acceptability | 16 (47%) |
| Satisfaction / meeting needs | 11 (32%) |
| Usefulness / helpfulness / effectiveness | 10 (29%) |
| Intention to use / actual use | 8 (24%) |
| Time / efficiency / duration | 7 (21%) |
| Comprehensibility / intelligibility | 5 (15%) |
| Objective outcome quality | 4 (12%) |
| Other | 16 (47%) |

**Table 4:** Quality indicators and the number of studies they were used in (including percentages)

| Judge | Number of studies |
|---|---|
| Provider | 18 (53%) |
| Recipient | 13 (38%) |
| Native speaker / bilingual | 8 (24%) |
| Translator / translation student | 3 (9%) |
| User | 3 (9%) |
| Other | 5 (15%) |

**Table 5:** Categories of judges and the number of studies they were used in (including percentages)

### 4.4 Overall performance

Of the 34 studies in the sample, 22 (65%) reported overall positive results with regard to the performance of the MT app(s) under study. 8 studies (24%) yielded mixed results, while only 4 studies (12%) were unequivocally negative in their final judgement. Mixed results mainly stemmed from differences between tested apps or variants of apps (e.g., Bouillon et al., 2017; Turner et al., 2019; Starlander et al., 2005) or different outcomes for different quality indicators (e.g., Seligman & Dillinger, 2015; Herrmann-Werner et al., 2021; Calefato et al., 2016).

Because of small cell sizes, the number of meaningful Chi-squared tests that we could run was limited. However, the outcomes of the tests that we did conduct show that an overall positive evaluation occurred more often than expected if the app was created by the authors themselves ($\chi^2(2) = 6.09$, $p < 0.05$) and if the test involved real-life situations ($\chi^2(2) = 7.55$, $p < 0.05$). Conversely, a negative overall evaluation occurred more often than expected if accuracy was used as a quality indicator ($\chi^2(2) = 7.32$, $p < 0.05$).

## 5 Conclusions and discussion

The aim of this study was to gain insight into (1) how the performance of real-time MT apps has been evaluated in previous research and (2) which overall picture concerning the performance of real-time MT apps emerges from the research conducted so far. To this end, we conducted a literature review in which we coded 34 published studies reporting an evaluation of real-time MT apps and their output.

Based on the results, we can conclude that the vast majority of studies have tested the app(s) during actual dialogues between interlocutors who did not speak each other's language (RQ1). In about half of those studies, a predefined scenario was used; in the other half, participants used the app(s) during their daily work. The most commonly used quality indicators were the perceived ease of use, the accuracy of the translations, the satisfaction with the user experience, and the perceived usefulness (RQ2). Therefore, it should not come as a surprise that users (both providers and recipients) were frequently employed as judges. Professional translators were involved in only a handful of studies (RQ3). Finally, 22 of the 34 studies came to a positive overall conclusion regarding the performance of the tested app(s). Only 4 studies reported mainly negative results (RQ4).

These outcomes suggest that fitness-for-purpose has indeed been an important guiding principle in previous studies that evaluated real-time MT apps. This is understandable, as many quality indicators used for the evaluation of written MT output are less applicable when MT is used to support synchronous dialogue. In addition, many studies were conducted with a view to a concrete professional context (e.g., communication between doctors and patients), which can explain why the focus was mainly on the course and the outcome of the dialogue as a whole, and less on the literal content of individual utterances within that dialogue.

At the same time, there are a number of observations that are cause for concern, both from a methodological as well as from a practical point of view. First of all, many studies are not clear about the definitions of their quality indicators, and even the most commonly used dependent variables are operationalized in many different ways. This not only reduces the comparability of studies, but also the possibility for professionals to make an evidence-based decision regarding the best app for their specific purpose. A similar point can be made with regard to the wide variety of language combinations examined and the lack of standardization in test scenarios. These methodological choices also add variance to the data that can obscure insight into the overall performance of the apps under investigation.

Another striking finding is that providers of care, education or services are asked about their experiences more often than recipients. One could argue that real-time MT apps are more likely to benefit recipients, as they can remove language barriers and increase the likelihood that recipients' wishes and concerns are well understood by providers. However, if a doctor or teacher feels that a dialogue that was supported by a real-time MT app has gone well, that does not necessarily mean that the other party involved in the dialogue has also experienced it that way. Therefore, it is advisable to always involve both parties in future evaluations.

Finally, only a few studies have attempted to establish objectively whether the translated dialogue also led to the desired outcome – in most cases, a correct diagnosis (e.g., Bouillon et al., 2017; Leite et al., 2016; Spechbach et al., 2019; Starlander et al., 2005). Although determining the correctness or objective desirability of an outcome is not possible in all professional situations, especially in contexts such as healthcare and education, one would expect that more attention would be devoted to what ultimately matters: A patient who recovers and a student who learns.

Of course, our study also has its limitations. Because reference lists played an important role in identifying potentially relevant studies, it is possible that we have overlooked previous research from certain professional domains. Since the majority of the studies in our sample (82%) were conducted in the context of healthcare, we could not compare the performance of real-time MT apps – nor the expectations of their users – across professional domains. In addition, some features of previous studies were not explicitly coded, such as the distinction between fixed-phrase translators and MT apps that can handle unrestricted input. Moreover, because the final sample was relatively small, we were only able to make a limited number of comparisons in our statistical analyses.

Therefore, we hope that future studies can investigate more systematically which variables explain the differences in performance between real-time MT apps. In addition, the various definitions and operationalizations of quality indicators can be mapped, so that more insight is gained into their interrelationships and conceptual overlap. Finally, it may be possible to develop and validate a more or less standardized test protocol that can increase the comparability of future studies.

### Acknowledgement

The author would like to thank Magda de Bruin for her contribution to the coding process.

## References

Bouillon, P., Gerlach, J., Spechbach, H., Tsourakis, N., & Halimi Mallem, I. S. (2017). Babeldr vs Google Translate: A user study at Geneva university hospitals (HUG). In *20th Annual Conference of the European Association for Machine Translation (EAMT)*.

Calefato, F., Lanubile, F., Conte, T., & Prikladnicki, R. (2016). Assessing the impact of real-time machine translation on multilingual meetings in global software projects. *Empirical Software Engineering, 21(3),* 1002-1034.

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to human and machine translation quality assessment. In J. Moorkens, S. Castilho, F. Gaspari and S. Doherty (Eds.), *Translation quality assessment*. Cham: Springer, pp. 9–38.

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

Creswell, J.W. (2014). *Research design: Qualitative, Quantitative, and Mixed Methods Approaches.* Los Angeles: Sage.

Directorate-General for Translation (2016). DGT guidelines for evaluation of outsourced translation. *Ares (2016) 3157529*.

Dorr B., Snover M. and Madnani N. (2011). Chapter 5.1 introduction. In Olive J., McCary J. and Christianson C. (Eds.), *Handbook of Natural Language Processing and Machine Translation.* DARPA Global Autonomous Language Exploitation. New York: Springer, pp. 801–803

Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research synthesis methods, 11(2),* 181-217.

Haith-Cooper, M. (2014). Mobile translators for non-English-speaking women accessing maternity services. *British Journal of Midwifery, 22(11),* 795-803.

Herrmann-Werner, A., Loda, T., Zipfel, S., Holderried, M., Holderried, F., & Erschens, R. (2021). Evaluation of a Language Translation App in an Undergraduate Medical Communication Course: Proof-of-Concept and Usability Study. *JMIR mHealth and uHealth, 9(12),* e31559.

Hwang, K., Williams, S., Zucchi, E., Chong, T. W., Mascitti-Meuter, M., LoGiudice, D., ... & Batchelor, F. (2022). Testing the use of translation apps to overcome everyday healthcare communication in Australian aged-care hospital wards—An exploratory study. *Nursing open.*

Jiménez-Crespo, M. A. (2018). Crowdsourcing and translation quality: Novel approaches in the language industry and translation studies. In J. Moorkens, S. Castilho, F. Gaspari and S. Doherty (Eds.), *Translation quality assessment*. Cham: Springer, pp. 69-93.

Lacruz I., Denkowski M. and Lavie A. (2014). Cognitive demand and cognitive effort in post-editing. In *Proceedings of the Third Workshop on Post-Editing Technology and Practice. 11th Conference of the Association for Machine Translation in the Americas, Vancouver, BC, Canada.*

Leite, F. O., Cochat, C., Salgado, H., da Costa, M. P., Queirós, M., Campos, O., & Carvalho, P. (2016). Using Google Translate© in the hospital: a case report. *Technology and Health Care*, *24(6),* 965-968.

Moorkens, J., Castilho, S., Gaspari, F. and Doherty, S. (Eds.) (2018). *Translation quality assessment*. Cham: Springer.

Panayiotou, A., Hwang, K., Williams, S., Chong, T. W., LoGiudice, D., Haralambous, B., ... & Batchelor, F. (2020). The perceptions of translation apps for everyday health care in healthcare workers and older people: A multi-method study. *Journal of Clinical Nursing*, *29(17-18),* 3516-3526.

Pitman, J. (2021, 28 April). Google Translate: One billion installs, one billion stories. Retrieved from https://blog.google/products/translate/one-billion-installs/ on 25 March 2022.

Przybocki M., Le A., Sanders G., Bronsart S., Strassel S. and Glenn M. (2011). Chapter 5.4.3 Post-editing. In Olive J., McCary J. and Christianson C. (Eds.), *Handbook of Natural Language Processing and Machine Translation*. DARPA Global Autonomous Language Exploitation. New York: Springer

Rivera-Trigueros, I. (2021). Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation*, 1-27.

Seligman, M. (1997). Interactive real-time translation via the Internet. *Working Notes, Natural Language Processing for the World Wide Web*, 24-26.

Seligman, M., & Dillinger, M. (2015). Evaluation and Revision of a Speech Translation System for Health Care. In *Proceedings of International Workshop for Spoken Language Translation 2015* (pp. 3-4).

Spechbach, H., Gerlach, J., Karker, S. M., Tsourakis, N., Combescure, C., & Bouillon, P. (2019). A speech-enabled fixed-phrase translator for emergency settings: Crossover study. *JMIR medical informatics*, *7(2)*, e13167.

Starlander, M., Bouillon, P., Rayner, M., Chatzichrisafis, N., Hockey, B. A., Isahara, H., ... & Santaholma, M. (2005). Breaking the language barrier: machine assisted diagnosis using the medical speech translator. *Studies in health technology and informatics*, *116*, 811-816.

Tao, U. (2022, 25 January). Top ten free translator apps 2022. Retrieved from https://www.timekettle.co/blogs/tips-and-tricks/top-10-free-translator-apps-2020 on 25 March 2022.

Turner, A. M., Choi, Y. K., Dew, K., Tsai, M. T., Bosold, A. L., Wu, S., ... & Meischke, H. (2019). Evaluating the usefulness of translation technologies for emergency response communication: A scenario-based study. *JMIR public health and surveillance*, *5(1)*, e11171.

Van Egdom, G. M. W., & Pluymaekers, M. (2019). Why go the extra mile? How different degrees of post-editing affect perceptions of texts, senders and products among end users. *Journal of specialised translation, 31*, 158-176.

Vieira, L. N., O’Hagan, M., & O’Sullivan, C. (2021). Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society*, 24(11), 1515-1532.

Yang, J., & Lange, E. D. (1998). SYSTRAN on AltaVista. A user study on real-time machine translation on the Internet. In *Conference of the Association for Machine Translation in the Americas* (pp. 275-285). Springer, Berlin, Heidelberg.

## Appendix A. Overview of studies included in the literature review.

Ahmed, A. N., Chit, S. C., & Omar, M. (2016). Design and evaluation of a multilanguage instant messaging application. *Proceedings of Knowledge Management International Conference 2016*, Chiang Mai, Thailand.

Aiken, M., & Park, M. (2020). A Comparison of two Multilingual Meeting Systems. *International Journal of Computer and Technology*, *20*, 38-44.

Albrecht, U. V., Behrends, M., Matthies, H. K., & von Jan, U. (2013). Usage of multilingual mobile translation applications in clinical settings. *JMIR mHealth and uHealth*, *1(1)*, e2268.

Beh, T. H. K., & Canty, D. J. (2015). English and Mandarin translation using Google Translate software for pre-anaesthetic consultation. *Anaesthesia and intensive care*, *43(6)*, 792.

Bouillon, P., Gerlach, J., Spechbach, H., Tsourakis, N., & Halimi Mallem, I. S. (2017). Babeldr vs Google Translate: A user study at Geneva university hospitals (HUG). In *20th Annual Conference of the European Association for Machine Translation (EAMT)*.

Calefato, F., Lanubile, F., Conte, T., & Prikladnicki, R. (2016). Assessing the impact of real-time machine translation on multilingual meetings in globalsoftware projects. *Empirical Software Engineering, 21(3)*, 1002-1034.

Chen, X., Acosta, S., & Barry, A. E. (2017). Machine or human? Evaluating the quality of a language translation mobile app for diabetes education material. *JMIR diabetes, 2(1)*, e13.

Day, K. J., & Song, N. (2017). Attitudes and concerns of doctors and nurses about using a translation application for in-hospital brief interactions with Korean patients. *BMJ Health & Care Informatics*, *24(3)*.

Ehsani, F., Kinzey, J., Zuber, E., Master, D., & Sudre, K. (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* (pp. 54-59).

Freyne, J., Bradford, D., Pocock, C., Silvera-Tawil, D., Harrap, K., & Brinkmann, S. (2018). Developing digital facilitation of assessments in the absence of an interpreter: participatory design and feasibility evaluation with allied health groups. *JMIR formative research*, *2(1)*, e8032.

Haith-Cooper, M. (2014). Mobile translators for non-English-speaking women accessing maternity services. *British Journal of Midwifery, 22(11)*, 795-803.

Herrmann-Werner, A., Loda, T., Zipfel, S., Holderried, M., Holderried, F., & Erschens, R. (2021). Evaluation of a Language Translation App in an Undergraduate Medical Communication Course: Proof-of-Concept and Usability Study. *JMIR mHealth and uHealth, 9(12)*, e31559.

Hwang, K., Williams, S., Zucchi, E., Chong, T. W., Mascitti-Meuter, M., LoGiudice, D., ... & Batchelor, F. (2022). Testing the use of translation apps to overcome everyday healthcare communication in Australian aged-care hospital wards—An exploratory study. *Nursing open.*

Janakiram, A. A., Gerlach, J., Vuadens-Lehmann, A., Bouillon, P., & Spechbach, H. (2020). User Satisfaction with a Speech-Enabled Translator in Emergency Settings. *Digital Personalized Health and Medicine*, 1421-1422.

Kaliyadan, F. & Sreekanth, G. (2010). The use of Google language tools as an interpretation aid in cross-cultural doctor–patient interaction: a pilot study. *Informatics in primary care*, *18(2)*, 141-43.

Kapoor, R., Truong, A. T., Vu, C. N., & Truong, D. T. (2020). Successful Verbal Communication Using Google Translate to Facilitate Awake Intubation of a Patient With a Language Barrier: A Case Report. *A&A Practice, 14(4)*, 106-108.

Leite, F. O., Cochat, C., Salgado, H., da Costa, M. P., Queirós, M., Campos, O., & Carvalho, P. (2016). Using Google Translate© in the hospital: a case report. *Technology and Health Care*, *24(6)*, 965-968.

Narang, B., Park, S. Y., Norrmén-Smith, I. O., Lange, M., Ocampo, A. J., Gany, F. M., & Diamond, L. C. (2019). The use of a mobile application to increase access to interpreters for cancer patients with limited English proficiency: a pilot study. *Medical care*, *57(Suppl 6 2)*, S184.

Ozaki, S., Matsunobe, T., Yoshino, T., & Shigeno, A. (2011). Design of a face-to-face multilingual communication system for a handheld device in the medical field. In *International Conference on Human-Computer Interaction* (pp. 378-386). Springer, Berlin, Heidelberg.

Panayiotou, A., Hwang, K., Williams, S., Chong, T. W., LoGiudice, D., Haralambous, B., ... & Batchelor, F. (2020). The perceptions of translation apps for everyday health care in healthcare workers and older people: A multi-method study. *Journal of Clinical Nursing*, *29(17-18)*, 3516-3526.

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

Ross, R. K., Lake, V. E., & Beisly, A. H. (2021). Preservice teachers' use of a translation app with dual language learners. *Journal of Digital Learning in Teacher Education*, *37(2)*, 86-98.

Şahin, M., & Duman, D. (2013). Multilingual Chat through Machine Translation: A Case of English-Russian. *Meta: journal des traducteurs/Meta: Translators' Journal*, *58(2)*, 397-410.

Seligman, M., & Dillinger, M. (2015). Evaluation and Revision of a Speech Translation System for Health Care. In *Proceedings of International Workshop for Spoken Language Translation 2015* (pp. 3-4).

Şentürk, E., Orhan-Sungur, M., & Özkan-Seyhan, T. (2021). Google Translate: Can It Be a Solution for Language Barrier in Neuraxial Anaesthesia? *Turkish Journal of Anaesthesiology and Reanimation*, *49(2)*, 181.

Silvera-Tawil, D., Pocock, C., Bradford, D., Donnell, A., Freyne, J., Harrap, K., & Brinkmann, S. (2021). Enabling Nurse-Patient Communication With a Mobile App: Controlled Pretest-Posttest Study With Nurses and Non–English-Speaking Patients. *JMIR nursing*, *4(3)*, e19709.

Soller, R. W., Chan, P., & Higa, A. (2012). Performance of a new speech translation device in translating verbal recommendations of medication action plans for patients with diabetes. *Journal of diabetes science and technology*, *6(4)*, 927-937.

Spechbach, H., Gerlach, J., Karker, S. M., Tsourakis, N., Combescure, C., & Bouillon, P. (2019). A speech-enabled fixed-phrase translator for emergency settings: Crossover study. *JMIR medical informatics*, *7(2)*, e13167.

Stankevičiūtė, G., Kasperavičienė, R., & Horbačauskienė, J. (2017). Issues in machine translation: a case of mobile apps in the Lithuanian and English language pair. *International journal on language, literature and culture in education*, *4*, 75-88.

Starlander, M., Bouillon, P., Flores, G., Rayner, M., & Tsourakis, N. (2008). Comparing two different bidirectional versions of the limited-domain medical spoken language translator MedSLT. In *Proceedings of the 12th Annual conference of the European Association for Machine Translation* (pp. 176-181).

Taicher, B. M., Alam, R. I., Berman, J., & Epstein, R. H. (2011). Design, implementation, and evaluation of a computerized system to communicate with patients with limited native language proficiency in the perioperative period. *Anesthesia & Analgesia*, *112(1)*, 106-112.

Turner, A. M., Choi, Y. K., Dew, K., Tsai, M. T., Bosold, A. L., Wu, S., ... & Meischke, H. (2019). Evaluating the usefulness of translation technologies for emergency response communication: A scenario-based study. *JMIR public health and surveillance*, *5(1)*, e11171.

Villalobos, O., Lynch, S., DeBlieck, C., & Summers, L. (2017). Utilization of a mobile app to assess psychiatric patients with limited English proficiency. *Hispanic Journal of Behavioral Sciences*, *39(3)*, 369-380.