Research article
Submitted: 4 Sept. 2021
Accepted: 20 Dec. 2021

ASSESSING SPEECH-TO-SPEECH TRANSLATION QUALITY: CASE STUDY OF THE ILA S2S APP

Marija Omazić and Martina Lekić

University of Osijek

Abstract

Machine translation (MT) is becoming qualitatively more successful and quantitatively more productive at an unprecedented pace. It is becoming a widespread solution to the challenges of a constantly rising demand for quick and affordable translations of both text and speech, causing disruption and adjustments of the translation practice and profession, but at the same time making multilingual communication easier than ever before. This paper focuses on the speech-to-speech (S2S) translation app Instant Language Assistant (ILA)[1], which brings together the state-of-the-art translation technology: automatic speech recognition, machine translation and text-to-speech synthesis, and allows for MT-mediated multilingual communication. The aim of the paper is to assess the quality of translations of conversational language produced by the S2S translation app ILA for en-de and en-hr language pairs. The research includes several levels of translation quality analysis: human translation quality assessment by translation experts using the Fluency/Adequacy Metrics, light-post editing, and automated MT evaluation (BLEU). Moreover, the translation output is assessed with respect to language pairs to get an insight into whether they affect the MT output quality and how. The results show a relatively high quality of translations produced by the S2S translation app ILA across all assessment models and a correlation between human and automated assessment results.

Keywords: speech translation technology, speech-to-speech translation apps, translation quality assessment, ILA

[1] The authors are grateful to the TranslateLive CEO Peter Hayes, who provided us with free access to the ILA app, thus making this study possible.

1. Introduction

In a constantly developing and globalized world, technological advances have brought a major shift in translation as a means of multilingual communication. Computer-assisted translation (CAT) tools and machine translation (MT) "have increased productivity and quality in translation, supported international communication, and demonstrated the growing need for innovative technological solutions to the age-old problem of the language barrier" (Doherty 2016: 947). Technological advancements have enabled the emergence of game-changing translation solutions for both text and speech, which are becoming widely accessible to the general public and translators alike.

One such innovative translation technology, which is slowly entering every domain of human life, is automated speech translation. Using software to translate text from one language into another in a matter of seconds is an already widely accepted feature of MT-mediated communication. However, speech-to-speech (S2S) translation is a far more complex undertaking, as it requires impressive state-of-the-art technology to produce the spoken output in the target language from the spoken input in the source language. "The task of translating acoustic speech signals into text in a foreign language is a complex and multi-faceted task that builds upon work in automatic speech recognition (ASR) and machine translation" (Sperber and Paulik 2020: 7409). In comparison to speech-to-text translation (S2T), speech-to-speech translation (S2S) also includes the conversion of the translated text into spoken language, i.e. text-to-speech synthesis (T2S). Owing to their multi-faceted design and workflow, speech translation apps encounter various difficulties in the translation process. Clearly, a "better ASR, MT, or T2S performance makes for better speech translation performance" (Waibel and Fügen 2008: 70). As the speech translation process entails tree separate steps, a minor mistake in the first one can become quite a serious one by the end of the process. The first challenge is automatic speech recognition, where speech is recognized and transcribed. In the second step, machine translation is performed, translating the source language text into the text in the target language. Lastly, text-to-speech synthesis creates the speech signal from the translated text (Arora et al. 2013: 209).

Speech translation solutions and applications are becoming more widely accessible to the general public, which raises the issues of output quality and usability as factors for their acceptance. The output quality of one such solution, the Instant Language Assistant (ILA), was tested in this case study.

2. Literature review

There is a large body of research on the translation quality assessment (TQA) of human and machine translation, both using human and automatic TQA methods; for a comprehensive overview see Castilho et al. (2018) and Moorkens et al. (2018). Among the most influential and widespread models of TQA assessment are the LISA QA model developed by the Localisation Industry Standards Association, which is based on error categorisation, error severity assessment and penalisation, and the Dynamic Quality Framework (DQF), a set of industry-developed tools for evaluating translation quality developed by TAUS (Translation Automation User Society), which considers variables such as communicative function, end-user requirements, context and translation mode. The Multidimensional Quality Metrics framework (MQM), on the other hand, is a shared quality metric for human translation (HT) and MT quality evaluation[2]. However, there is still no consensus on what the best model may be.

Here we focus on one of the most frequently used models for *human TQA* is the Adequacy-Fluency Metrics for evaluating MT quality (Koehn 2009: 218), a two-dimensional evaluation metric used by human evaluators or assessors, aiming to "provide a more balanced view on translation quality" (Banchs et al. 2015: 472). The Linguistic Data Consortium (2005) defined translation adequacy as a response to the question "How much of the meaning expressed in the gold-standard translation or the source is also expressed in the target translation?". Fluency is defined as a response to the question to what extent the translation is "one that is well-formed grammatically, contains correct spellings, adheres to common use of terms, titles and names, is intuitively acceptable and can be sensibly interpreted by a native speaker". Translations are rated by human assessors on a five-point scale to indicate how fluent (from flawless to incomprehensible) and how accurate they

[2] For a detailed overview of MQM, terms, definitions and 19 quality issues in the MQM core see http://www.qt21.eu/mqm-definition/definition-2015-06-16.html

are (from all to no original meaning preserved). The Adequacy-Fluency Metrics with a fine-grained error taxonomy was proposed by Daems and Macken (2013) and Daems et al. (2014) for the English-Dutch language pair to help human evaluators annotate the errors found in target texts. Their categories of errors relating to adequacy and fluency are shown in Table 1.

Table 1. Adequacy-Fluency Metrics (Daems and Macken 2013)

ADEQUACY	**FLUENCY**				
Meaning shift	**Grammar and Syntax**	**Lexicon**	**Spelling and Typos**	**Style and Register**	**Coherence**
contradiction	article	wrong preposition	capitalization	register	conjunction
word sense disambiguation	comparative/ superlative	wrong collocation	spelling mistake	untranslated	missing info
hyponymy hyperonymy	singular/plural	non-existent word	compound	repetition	logical problem
terminology	verb form		punctuation	disfluent sentence/construction	paragraph
quantity	article-noun agreement		typo	short sentence	inconsistency
time	noun-adjective agreement			long sentence	other
meaning shift caused by punctuation	subject-verb agreement			text type	
meaning shift caused by misplaced word	reference			other	
deletion	missing constituent/preposition				
addition	word order				
explicitation	structure – other				
coherence	grammar – other				
inconsistent terminology					
other					

For a more detailed overview of the error categories, their definitions, examples and annotation see Daems and Macken (2013).

Most of the *automated TQA* metrics evaluate the translated content based on a similarity between the MT output and a reference translation. The most widely used automated TQA metrics are BLEU, NIST, METEOR, TER, or CharacTER. The BLEU

(Bi-Lingual Evaluation Understudy) Metric was proposed by Papineni et al. (2002) and despite its flaws, it is the most favoured measure used today. BLEU measures the correspondence between the MT output and a reference translation or translations and measures direct word-to-word and word string cluster similarity. Since it works on a direct similarity-based method, BLEU does not measure the overall quality or the accuracy of a translation, but measures adequacy of MT by comparing the word precision and brevity of MT compared to the reference translation. The BLEU score is given on a scale of 0 to 100, with 100 representing a 100% match of the MT translation output with the reference translation. The creators of BLEU highlight the following advantage: "BLEU's strength is that it correlates highly with human judgments by averaging out individual sentence judgment errors over a test corpus rather than attempting to divine the exact human judgment for every sentence: quantity leads to quality" (Papineni et al. 2002: 318), which has been disputed in many subsequent studies (e.g. Callison-Burch et al. 2006: 255). However, as BLEU is based on the degree of similarity, a correct translation worded differently will receive a low score, which is a major shortcoming of BLEU.

As the automatic translation of speech is also machine translation, the same methods used for the TQA of MT can be applied to assess speech translation quality, with a few important caveats. For a comprehensive TQA of speech, the evaluation models need to be expanded to accommodate for the assessment of speech recognition quality and speech synthesis quality to achieve a comprehensive end-to-end evaluation. One such attempt was the evaluation of the S2S translation system in the TC-STAR project[3], performed by Hamon, Mostefa and Choukri (2007), who assessed three-minute segments of European Parliament plenary sessions and broadcast news recordings translated by automatic systems and human interpreters. The speech recognition quality was measured using word error rate (WER), BLEU was used as the automatic metric for MT of speech, the human TQA was carried out using the fluency and adequacy metrics, and the text-to-speech evaluation was made using a test to evaluate the prosody, expressiveness and voice conversion. The authors found that human interpreters obtained better results on fluency, whereas the automatic systems performed better on accuracy.

[3] The aim of the TC-STAR project was to advance research in all core S2S technologies. See: http://tcstar.org/

Le, Lecouteux and Besacier (2018) experimented with using joint ASR and MT features for automatic quality assessment of speech translation and gave more weight to MT assessment, with ASR assessment bringing only complementary information. Chen et al. (2017) evaluated the quality of a language translation mobile app *iTranslate* output compared with human interpreters using a five-point scale to assess fluency, adequacy, meaning and severity, and found that *iTranslate* generally provided translation accuracy comparable to human translators on simple sentences, but made more errors when translating difficult sentences. Müller et al. (2016a, 2016b) measured the perceived quality of the KIT lecture translation system for university lectures in German using exit polls, short surveys and large questionnaires evaluating ASR and MT, and standardised interviews to get more user impressions and suggestions.

When it comes to translating bilingual dialogues, there are a few relevant projects and studies. The Vermobil project (Wahlster 2000) is a bidirectional mobile S2S translation system for spontaneous dialogues in business-like domains for German, English and Japanese. The success rate of this dialogue interpretation service was estimated at 75% for word recognition rate, 80% of approximately correct translations that preserve the speaker's intended effect, and 90% success rate for dialogue tasks in end-to-end evaluations with real users. Another corpus that contains MT-mediated interaction is the Field Experiment Data by Takezawa et al. (2007), which also provides human judgement of overall translation quality, but not at the sentence level. The Microsoft Speech Language Translator (MSLT) corpus (Federmann and Lewis 2016) was created to evaluate end-to-end conversational speech translation quality of actual Skype conversations powered by Microsoft's MT engine Microsoft Translator[4] in English, French and German. A recent paper by Bawden et al. (2021) presents an English-French corpus of bilingual spontaneous written dialogues for machine translation, mediated by neural MT systems and evaluated by human judges, using very simple three-point scales (perfect, medium, poor) and a simple error categorisation (grammar, meaning, style, word choice, coherence and other). Their results show that the quality of MT output depends on the language pair, as translations into English scored better than translations into

[4] The corpus is available at https://msropendata.com/datasets/54813518-4ea6-4c39-9bb2-b0d1e5f0c187.

French. However, this type of research of MT-mediated multilingual communication is still rare and needs more concentrated effort focusing on different types of content and different media and platforms which provide MT-assisted interaction. In her MA thesis, Lekić (2021) assessed selected speech translation apps for the en-de-hr combinations, which served as a starting point for the research presented in this paper.

3. Quality Assessment of ILA-produced Translations

3.1 Research aims

The aim of this study is to test the Instant Language Assistant (ILA) and assess the quality of machine translations of conversational language in everyday situations, such as simple conversations in the bank, at the hotel and in the store, using several different translation quality assessment metrics and to compare their results. As the quality of the machine-translated output depends on the language pair, with languages of lesser diffusion and fewer resources and heavily inflected languages usually producing lower quality MT output, this paper also aims to address the issue of quality in relation to the language pair in which the test dialogues are conducted (en-de *vs.* en-hr).

3.2. Research Design

Our research was conducted using three bilingual dialogues taking place in different real-life settings – a bank, a hotel and a grocery store, each performed in English-German and English-Croatian language combinations. Previous research mostly looked at uninterrupted speech in one language, and its machine translation into another language, such as parliamentary sessions, news reports, multilingual lectures or diabetes instructions (Hamon, Mostefa and Choukri 2007; Müller et al. 2016a, 2016b; Chen et al. 2017), and only a few very recent studies included bilingual dialogues (Bawden et al. 2021, Federmann and Lewis 2016). The dialogues we used were scripted for the purpose of this study, but the speakers were instructed not to adjust the speed or accommodate for potential ASR errors that might impact the MT quality. The dialogues were short every-day exchanges between two interlocutors. Each dialogue had 250 words on average (500 words if

we count in the translation), with about 20 short exchanges between speakers, lasting about 2 minutes. The duration is longer than the word count may suggest, because it took twice the time for both input in the source and output in the target language of each speaking turn. The bilingual dialogues were translated by the ILA app. Each translation underwent a detailed human translation quality assessment of fluency and adequacy, conducted by three independent evaluators with 10 to 25 years of professional translation experience. Their quality assessment was based on the Adequacy-Fluency Metrics for human evaluation of MT (Daems and Macken 2013). Next, the ILA-produced translations were lightly post-edited. Light post-editing was chosen because the purpose of the whole exercise was to enable communication and understanding between interlocutors speaking two different languages, and not to get a perfect, publishable translation. In addition, the ILA-produced translations were assessed using the Automated Translation Metrics, providing an objective, automatic quality assessment of the MT output. We used the post-edited translations as reference translation. Lastly, the results of all translation quality assessment methods were compared with regards to language combination.

3.2.1. About the Instant Language Assistant

The app we tested, ILA, is an S2S app designed to perform the role of a language mediator between people who speak different languages but need to communicate with each other.

Figure 1. The ILA device

The initial idea of the TranslateLive App was introduced in 2017, and the first ILA prototype was launched in 2019. TranslateLive has created two devices specialized for specific usage situations: The ILA Traveller and the ILA Pro device. ILA Traveller is small and compact and it can accompany users on their journeys. The ILA Pro is a larger model with two screens, designed to be a stationary desk device. Businesses, the Government, emergency services, hostels, airports, concierge desks, and many more can benefit from having a translation device at the reach of their hand. To start translating, a user just presses the button and starts speaking in their preferred language. The speech is converted to text format, so the speaker can ensure everything is recognized and transcribed correctly. The translated text is displayed to the person on the other end in their selected language, with the option to be read out aloud. Then, the second person presses the button to respond.

The TranslateLive platform[5] uses several features to drive its live translation such as the Compute Engine, Translation API, and Cloud Speech-to-Text. The Google Compute Engine "delivers virtual machines running in Google's innovative data centres and worldwide fibre network"[6], providing the necessary performance for real-time translation without latency issues. The Translation API is an instant translator for websites and apps. The Speech-to-Text Cloud converts speech into text format using an API powered by Google's AI technologies[7]. Live automated language translation uses third-party providers like Google, Microsoft, Amazon and Apptek for non-specifically trained customer systems.

One of the most important features of the ILA app is its accessibility. The TranslateLive app is available for both iOS and Android, but it can also be used by anyone with just a web browser. This is important because ILA is not only easy to use but also easy to access at any time and from anywhere. Secondly, ILA is instant and accurate. It enables real-life conversations with minimum delay as the speech-to-text feature is constantly improving and upgrading for better-quality translations. Thirdly, ILA is suitable for people with disabilities, such as the deaf, blind and hard of hearing, as it gives them the possibility to speak or write what

[5] https://www.translatelive.com/ accessed on 21 June 2021.

[6] https://console.cloud.google.com/marketplace/details/google-cloud-platform/compute-engine?pli=1 accessed on 21 June 2021.

[7] https://cloud.google.com/speech-to-text/ accessed on 21 June 2021.

they want to say as well as to read and hear the translation, thus making it possible for them to successfully communicate with anyone. All the conversations are encrypted, private and HIPAA compliant. Last but not the least, the fact that speakers can check the speech on their own screen, ensures a much higher accuracy rate, which is important for a successful communicative act[8].

Finally, ILA is customizable, i.e. you can pre-load the phrases you need for better accuracy. Currently, the app is being trained based on specific strings to facilitate the vaccination process for both medical workers and patients who do not speak a common language. Questions such as *Is this your first dose of COVID-19 vaccine?* and phrases such as *Your arm may be sore for a day or two* have been translated into six languages, with many more in preparation. The App is using ASR and MT models which have been customized for Covid-related contect. We have translated Covid-related strings into Croatian as part of our collaboration in this project.

3.3. Human Translation Quality Assessment Methodology

The model used for the human TQA in this study is the Adequacy-Fluency Metrics for evaluating MT quality proposed by Daems and Macken (2013). Even though it is a metric for evaluating text-to-text translation, in this study we used it to evaluate the translation output of the ILA app. To obtain relevant results, the Adequacy-Fluency Metrics categories shown in Table 1 above had to be slightly adjusted to the specificities of speech translation output. Therefore, error categories such as punctuation and initial sentence capitalisation were not taken into account since these are caused by ASR segmentation errors (i.e. the inability of ASR systems to break down input audio into sentence-like units) and do not constitute a translation error as such. However, as regards translations into German, noun capitalisation was considered to be an indicator of translation quality, as nouns are always capitalised in German, even though it is irrelevant for speech. We decided to do this because ILA also shows you a written transcript. When evaluating fluency, the evaluators only had access to the translation output, and not to the source speech. They focused on the fluency of the produced target language translation and answered the following question: *"Is the language in the output fluent?"*, regardless

[8] https://www.translatelive.com/ila-solutions/ accessed on 21 June 2021.

of the correct meaning. By contrast, when evaluating adequacy, they had access to both the source and the target texts and were able to compare them and answer the question: *"How much meaning has been preserved?".* In other words, they focused on how faithfully the information given in the source text had been translated into the target language, even if the translation output was not fully fluent.

Each evaluator was provided with the translation input, the MT output and a table with error types for each of the evaluated categories[9]. Starting with fluency, the evaluators identified and annotated each error found in the target text that affected the fluency of the translation. They did the same when evaluating adequacy, annotating all adequacy errors found when comparing the source and target texts. After annotating the errors, the evaluators graded the overall translation quality based on the number of fluency and adequacy errors for each of the dialogues using the five-point grading scale as suggested by the LDC (2005). As the dialogues were short, we did not normalize the results but used the absolute number of errors in each speech. Even though the assessors had very explicit annotation guidelines, they reported that it was sometimes difficult to classify the errors and provide the overall grade for the quality of translations, which has also been reported in previous research (Koehn and Monz 2006).

Table 2. Fluency-Adequacy Metrics Grading Scale

Grade	Translation fluency	Translation accuracy
excellent (5)	flawless	all meaning preserved
very good (4)	good quality	most meaning preserved
good (3)	non-native	much meaning preserved
sufficient (2)	disfluent	little meaning preserved
insufficient (1)	incomprehensible	no meaning preserved

[9] See Daems and Macken (2013) for an extensive overview of error types and annotation in the Fluency/Adequacy Metrics.

3.4. Quality Assessment Results

The results of the quality assessment are given separately for each of the two language pairs, for both adequacy and fluency. The number of errors presented here is absolute and not weighted as our dialogues were short (500 words if you count in both the input and the translated output). Average grades for the translation quality of en-de dialogues are presented in Table 3 and range from excellent (5) to good (3), both for fluency and adequacy. The translations for this language pair have been assessed as excellent, in other words all of the meaning was preserved and the outcome had no fluency issues. Even though we did not measure interrater agreement specifically, the results indicate a high degree of agreement between grades given by different assessors, which is not always the case with human evaluators.

Table 3. Fluency and Adequacy Assessment results for the English-German language pair

FLUENCY	**Evaluator 1**		**Evaluator 2**		**Evaluator 3**		
	Errors	Grade	Errors	Grade	Errors	Grade	**Average grade**
At the bank	4	**5**	6	**4**	3	**5**	**5**
At the hotel	4	**5**	10	**4**	5	**5**	**5**
At the store	9	**4**	13	**3**	12	**3**	**3**
ADEQUACY							
At the bank	3	**5**	6	**4**	4	**5**	**5**
At the hotel	4	**5**	4	**5**	5	**5**	**5**
At the store	9	**4**	9	**4**	12	**3**	**4**

Grades for the translation quality of en-hr dialogues (Table 4) range from excellent (5) to good (3) as well, but the averages for both fluency and accuracy are lower (4) than for en-de dialogues (5). The evaluators assessed the en-hr translations as being of good quality, with most original meaning preserved. The agreement between grades given by different evaluators is even higher in this case. This is rare as human evaluation is always subjective to a certain degree, which leads to interrater disagreement and the need to always have at least two evaluators, if not more in case of large disparities.

Table 4. Fluency and Adequacy Assessment results for the English-Croatian language pair

FLUENCY	**Evaluator 1**		**Evaluator 2**		**Evaluator 3**		
	Errors	Grade	Errors	Grade	Errors	Grade	**Average grade**
At the bank	8	**4**	10	**4**	4	**5**	**4**
At the hotel	5	**5**	10	**4**	9	**4**	**4**
At the store	9	**4**	12	**3**	12	**3**	**3**
ADEQUACY							
At the bank	6	**4**	12	**3**	6	**4**	**4**
At the hotel	12	**3**	5	**5**	8	**4**	**4**
At the store	6	**4**	6	**4**	9	**4**	**4**

3.5. Post-editing Results

In this section, each of the dialogues will be discussed separately. A table with the source text, the ILA-produced MT and the post-edited version is provided for each dialogue (only the post-edited sections are shown), as well as for each of the two language combinations. As the purpose of the ILA app is to ensure that speakers understand each other, without claiming that the output is 100% accurate, only light post-editing was done, not taking into account punctuation and other style errors which do not influence the comprehensibility of the translation, based on human judgement. According to TAUS (2016), full post-editing aims to achieve the quality similar to human translation and revision, or publishable quality, whereas light post-editing should reach a good enough quality.

Table 5. At the bank – English and German

Source text	**Machine translation**	**Post-editing**
Wie viel Geld brauche ich um die Kontos zur eröffnen?	how much money do I need to open the account	how much money do I need to open the accounts
Sure, let me do that for you now.	Lass mich das jetzt sicher für dich tun	Sicher Lassen Sie mich das jetzt ~~sicher~~ für ~~dich~~ Sie tun
Here you go. Now you have a checking and a savings account with a €250 deposit on each.	bitte schön, jetzt haben Sie ein Scheck- und ein Sparkonto mit Einzahl von jeweils 250 €	bitte schön, jetzt haben Sie ein ~~Scheck-~~ Giro- und ein Sparkonto mit Einzahlung von jeweils 250 €

The example marked in grey in Table 5 is the only post-edited section translated from German into English. The source text contains the plural form

Kontos, and the translation is *account*, singular. The other two examples where post-editing was necessary are translations from English into German. The sentence *Sure, let me do that for you now*, caused some problems to machine translation. Firstly, since the dialogue is held at the bank, the bank official should address the customer formally, which means that *you* should be translated as *Sie*. In addition, the adverb *sure* is misplaced in translation. In the last example, the more appropriate translation of the *checking account* would be *Giroaccount*, and *deposit* was mistranslated with a non-existent word **Einzahl* instead of *Einzahlung*.

Table 6. At the bank – English and Croatian

Source text	Machine translation	Post-editing
Koliko mi je novca potrebno za otvaranje računa?	how much money do I need to open an account	how much money do I need to open ~~an~~ the accounts
Sure, let me do that for you now.	sigurno mi to dopustite da	~~sigurno~~ naravno ~~mi to dopustite da~~ sada ću Vam to učiniti
Here you go. Now you have a checking and a savings account with a €250 deposit on each.	evo, sada imate ček na štednom računu na kojem je uplaćeno 250 € depozita za svaku	evo, sada imate ~~ček na štednom~~ tekući i štedni račun~~u~~ ~~na kojem~~ i na svaki je uplaćeno 250 € depozita ~~za svaku~~
Puno hvala gospodine! Vrlo ste susretljivi!	thank you very much sir you are very accommodating	thank you very much sir you are very ~~accommodating~~ helpful

In the same dialogue but in the en-hr language combination (Table 6), four sections were post-edited. Two of the four sections are translations from Croatian into English, and they are marked in grey. The first example shows the same error as the translation from German into English in the previous table. However, in isolation, this is not an error at all, because the Croatian word *računa* has the same Genitive form in singular and plural. From the context it becomes clear that it refers to two types of accounts. This illustrates that machine translation functions well within sentence boundaries, but not beyond them, as it does not take the surrounding context into consideration and therefore could not recognize that this was a plural noun. The second example posed serious problems for the machine since the translation is neither correct nor complete. *Let me do that for you* was translated literally in part and the rest of the translation was simply left out. MT also had some problems with the next example. The most serious error is the

mistranslation of *a checking and a savings account*[10]. The translation mentions checks and only a savings account, when it should say *tekući i štedni račun*. In the last example, a more appropriate translation for *susretljiv* would be *helpful*, rather than *accommodating*.

Table 7. At the hotel – English and German

Source text	Machine translation	Post-editing
Wir hätten gerne ein Doppelzimmer mit Bad.	Hello, we would like a double room with a bathroom	Hello, we would like a double room with a ~~bathroom~~ bath
How long would you like to stay?	wie lange möchtest du bleiben	wie lange möchte~~st~~n ~~du~~ Sie bleiben

In the dialogue *At the hotel*, in the en-de language combination (Table 7), only two minor mistakes occur. The first is a translation from German into English, where *Bad* is translated as *bathroom* instead of *bath(tub)*. This is another illustration that MT operates within sentence boundaries, because as the dialogue progresses, it becomes clear that the only room that the concierge can offer is the one with a shower and not a bath(tub). The second error is connected to the grammatical number. Since two people came to the hotel, the concierge should address them formally as *Sie*, not *du*.

Table 8. At the hotel – English and Croatian

Source text	Machine translation	Post-editing
All right, let me check what is available.	u redu da provjerim što je dostupno	u redu samo da provjerim što je dostupno
How long would you like to stay?	koliko bi želio ostati	koliko biste ~~želio~~ željeli ostati
May I see your ID please, sir?	izvolite, vidim vašu osobnu iskaznicu, molim vas gospodine	~~izvolite~~ mogu li ~~vidim~~ vidjeti vašu osobnu iskaznicu, molim vas gospodine
If you need anything, just dial 0 on your room phone.	ako trebate bilo što, samo nazovite nulu na sobnom telefonu	ako trebate bilo što, samo ~~nazovite~~ birajte nulu na sobnom telefonu

[10] In the source speech we used the term *checking account*, which is a EN-US term, rather than the term *current account*, which would have resulted in a more accurate translation, because we did not want to oversimplify the input by avoiding potential pitfalls for MT.

In this dialogue, the post-editing was needed only in the segments translated from English into Croatian (Table 8). In the first example, the post-editor chose to correct the style to make the sentence sound more native-like. This was not essential, as this is an error of style which does not distort the meaning. In the second example, the post-editor had to change the number, just like in the German-English dialogues. The third example had a major mistranslation that had to be corrected (*May I* was translated as *izvolite* (here you are), and the rest was only lightly post-edited, without changing the sequence of the phrases. In Croatian it would have been more natural to start with the polite address *Gospodine, molim Vas mogu li…*. The fourth error is a mistranslation of the collocation *dial 0*, which should read *birajte nulu* in Croatian.

Table 9. At the store – English and German

Source text	Machine translation	Post-editing
Ich hätte gerne zwei Dutzend Eier und eine Flasche Milch.	I need two packets of eggs and one milk	I need two packets of eggs[11] and one bottle of milk
Außerdem möchte ich noch zwei Scheiben Räucherschinken	in addition you can slice two slices of smoked ham	in addition ~~you can~~ I would like to have ~~slice~~ two slices of smoked ham
That's £25. Here is your receipt.	das ist 25 £ hier ist Ihre Quittung	das ~~ist~~ wäre 25 £ hier ist Ihre ~~Quittung~~ Rechnung

In the dialogue *At the store* (Table 9), two out of three errors were made in translations from German into English. One intervention was needed in the translation from English into German. In the first example, MT used an uncountable noun *milk* as countable, and it had to be post-edited using a partitive noun *bottle of milk*. In the next example, the pleasant inquiry *ich möchte* was left out, so the post-editor added *I would like to have*. In the third example, the wrong verb form was used in German, namely, it should read *das wäre 25 £*. Also, the word *receipt*

[11] This is not a perfect collocation in English, it should be *two dozen eggs* or *24 eggs*, but as the translation was only lightly post-edited, minor departures from the original style were not post-edited.

was mistranslated as *Quittung,* which does not fit the register and should be *Rechnung*.

In the dialogue *At the store* led in English and Croatian (Table 10), the light post-editing was needed in four segments. Two of the errors appear in each of the language combinations. The first segment is the exact same as the one in the dialogue in English and German, and the MT result is exactly the same. The translation of the second segment has a collocation error because *cube sugar* should be *šećer u kocki* in Croatian. The next example needed to be post-edited because it is unnatural to say *imam to ovdje* in Croatian. In the last example, MT had the same problems translating the sentence from German and from Croatian as well.

Table 10. At the store – English and Croatian

Source text	**Machine translation**	**Post-editing**
Trebam dva paketa jaja i litru mlijeka.	I need two packets of eggs and one milk	I need two packets of eggs and one bottle of milk
Which sugar? Cube or Caster Sugar?	koji šećer kocka ili kristalni	koji šećer u ~~kocka~~ kocki ili kristalni
Yes I have, it's right here in the detergent department.	da imam to ovdje u odjelu deterdženta	da, ~~imam to~~ ovdje je ~~u~~ na odjelu deterdženta
Osim toga, može još dvije šnite dimljene šunke	in addition you can slice two slices of smoked ham	in addition I would like to have ~~you can slice~~ two slices of smoked ham

3.6. Automated Translation Metrics – BLEU Scores

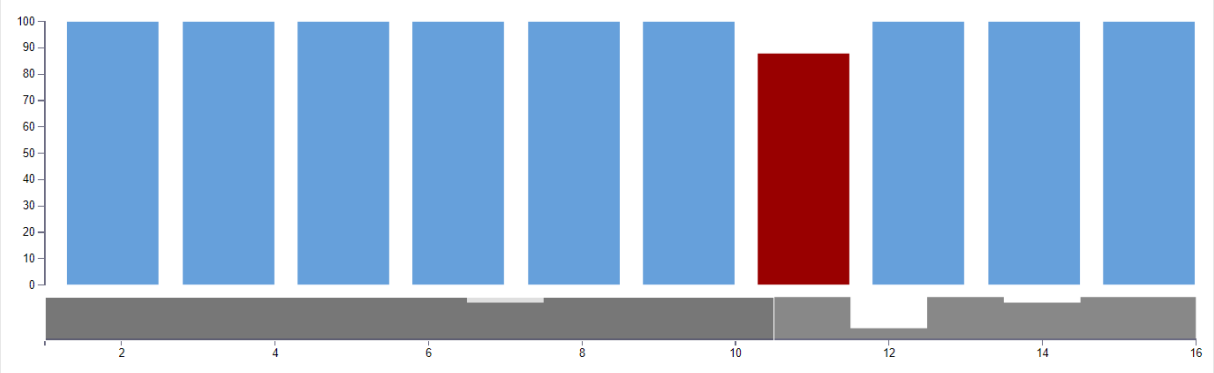
The automated metric we used for evaluating ILA-produced MT in our study is BLEU, as it has often been reported to correlate well with human translation quality assessment results. In this section, the results of the BLEU metrics of the ILA-produced MT output will be discussed. In the previous step, each of the translated dialogues was lightly post-edited. The post-edited translations served as the reference translation for this purpose. BLEU scores are presented in Table 11.

Table 11. Overview of BLEU scores

Dialogue	Language combination	Precision x Brevity[12]	BLEU score
At the bank	EN-DE	93.18 x 100.0	93.18
	EN-HR	82.23 x 97.35	80.06
At the hotel	EN-DE	96.32 x 100.00	96.32
	EN-HR	90.13 x 99.12	89.34
At the store	EN-DE	81.94 x 97.96	80.27
	EN-HR	79.62 x 98.99	78.82

It should be noted that for each of the dialogues, the language pair English and German received higher BLEU scores. The average score of all three dialogues in the language combination English-German is 89.92/100. For the language combination English-Croatian, the average score is 82.74/100. As expected, the English-Croatian dialogues scored lower, but the difference is surprisingly small. We must also point out that none of the ILA-produced translations scored lower than 70, which is often taken as the BLEU score threshold, so this could be taken as a relatively successful MT output. However, these results should be interpreted in the light of the fact that we used lightly post-edited translations as reference translations. If we had opted for full post-editing or for the human translation as reference texts, the scores would have been lower and more realistic. Another reason why BLEU scores are high is that the exchanges in our dialogues are short (8 to 12 words per exchange on average), and short sentences have been found to receive higher BLEU scores than long ones (cf. Chen et al. 2017 who evaluated the *iTranslation* app). Also, we had a small sample of less than 1000 sentences, which may affect the relevance of the score. One also needs to keep in mind that a correct translation can get a lower BLEU score just because a different word (a synonym) or phrase was used, but as we conducted only light post-editing this was not the case in our study. The BLEU metric will also score longer word strings better than shorter ones, even if shorter ones are also correct. These limitations of our study and of the BLEU metric need to be taken into account when interpreting our scores.

[12] Precision score takes into account adequacy and fluency (adequacy is satisfied if the same words are used in a translation as in a reference; fluency is satisfied if there are longer n-gram (word string) matches in translation and in a reference. Brevity penalty penalizes too short translations. High brevity scores match reference translations in length, word choice and word order.



<< < 7 > >> From 1 to 10 of all **16** sentences.

Sentence 7	BLEU	Length ratio	Text
Human	100.00	1.00	how much money do I need to open the accounts
Machine	88.01	1.00	how much money do I need to open the account

Figure 2. BLEU comparing sentences

One example of BLEU calculation at sentence level using our corpus example is shown in Figure 2. The translation is as long as the human reference translation, so it was not penalised for brevity, but it was penalised for lexical imprecision at the single word level, which is why it received a score of 88.01. The grammatical number error (*accounts vs. account*) was penalized by subtracting 12 points from the ideal score of 100. This is still a very high BLEU score.

3.7. Comparing Research Results

In this section, the results of all three assessments methods will be compared based on the language pair (Table 12). The aim of this comparison is to bring the results of the human fluency/adequacy TQA, light post-editing and automated machine translation metrics together to see whether there is agreement between them and establish whether the language combination affects the quality of the MT output across assessment methods.

When comparing the results of the quality assessment conducted by independent evaluators, the average grade was calculated from the grades for fluency and adequacy. The en-de dialogues scored one grade higher for both fluency and accuracy compared to en-hr dialogues. For the purpose of displaying the post-editing results, the number of necessary edits was counted, meaning that the higher the number of edits, the lower the MT output quality. The en-de

dialogues in had an average of 4.66 edits per dialogue. The en-hr dialogues needed more post-editing effort and had 8.33 edits per dialogue. The average BLEU score for the en-de dialogues was 89.91, whereas the en-hr dialogues obtained a slightly lower average score of 82.74.

Table 12. Comparing results with regards to language pair

Language pair		***EN-DE***	***EN-HR***
	Dialogue		
1. QUALITY ASSESSMENT of fluency/accuracy Excellent (5) to insufficient (1)	1	excellent (5)	very good (4)
	2	excellent (5)	very good (4)
	3	very good (4)	very good (4)
POST-EDITING Number of interventions > MT quality	1	4	9
	2	2	6
	3	8	10
AUTOMATED TRANSLATION METRICS BLEU score < 100	1	93.13	80.06
	2	96.32	89.34
	3	80.27	78.82

When it comes to the difference in the quality of ILA-produced translations in English and German and English and Croatian, in each round of assessment the en-de dialogues proved to be more successful across all three quality assessment metrics. There is a clear correspondence between human translation quality assessment of adequacy and fluency, the number of post-editing interventions and BLEU scores, i.e. higher quality translations had fewer errors, required fewer edits and had higher BLEU scores.

4. Conclusion

Professional translators, interpreters and users of translation will all agree that with the enormous technological advancements, MT has more and more advantages. Still, there are many aspects in which MT falls short. In this paper, a multi-layered research of the quality of the translation output produced by the speech-to-speech translation app ILA was conducted. Starting with a human translation quality assessment of fluency and adequacy, moving on to light post-editing and automated translation metrics, this research tried to encompass several measurable components of a successful translation and assess the overall quality of the S2S translations.

In our study, we focused on the quality assessment of machine translations of speech, but the quality of the automatic speech recognition and text to speech

synthesis technology used by the ILA app also need to be evaluated if one wants to get a clear picture of the overall quality of the app performance. Furthermore, these assessments need to be taken into account when developing a comprehensive quality assessment model for S2S translation. Although we have not measured it directly, our impression is that automatic speech recognition was satisfactory. With a similar number of errors in recognising speech in all of the three languages used, ILA converted speech to text well, regardless of the language, bearing in mind, however, that our sample was small. The users have to slightly adjust to the ASR by speaking loud enough and as clearly as possible, but when those prerequisites are met, ILA picks up the spoken input very well. If the speech is not converted to text well, the dialogue partners can repeat the wrong phrase to avoid mistranslations and misunderstandings. Also, the text-to-speech Synthesis technology produces an audio output of good quality. However, ILA does sound robotic, especially in Croatian, which is a feature that certainly needs to be taken into account when assessing the overall output quality and user satisfaction with the ILA app.

When it comes to the overall quality assessment of the dialogue translations produced by ILA, the translations were graded based on the fluency-adequacy translation metrics. The average grades of the two levels range from excellent (5), indicating a fluent translation that retained all of the original meaning, to very good (4), indicating a good quality translation that preserved most of the original meaning. In the translation post-editing process, between ten and two minor interventions per translation output were necessary to adapt the text according to grammatical and structural rules and avoid any misunderstandings between the dialogue participants. The automated translation metrics assessed the translations with the minimum BLEU score of 78.82, and a maximum of 96.32/100. Interestingly, the human evaluation of the translation output matches BLEU scores and the post-editing effort to a great extent, thereby confirming the accuracy and conformity of all three quality assessment techniques.

Overall, the research results indicate a relatively high quality of translations produced by the S2S app ILA. The main advantages are terminological precision and grammatical correctness as some of the most important preconditions for a successful translation. The main disadvantage is the inability of the MT to

"understand" communicative acts of illocution, politeness, and implied references in conversation. In other words, "MT systems suffer from not being able to anticipate context like human interpreters" (Müller et al. 2016a: 83). When it comes to variations in the quality of the ILA-produced translations based on the language pair in which the dialogues were led and translated into, all of the assessment methods show that the translation output in the language pair English-German was scored better than the translation output in English-Croatian. Still, with the average grade of very good (4) for both adequacy and fluency, the translations in English-Croatian were assessed as being of good quality and retaining most of the meaning, not far behind those in English-German graded as excellent (5).

Further research on S2S apps could conduct a detailed analysis of all the three layers of language technology necessary for the production of S2S translation. On the level of ASR, the Word Error Rate (WER), a common metric used to measure the performance of speech recognition could be applied. When it comes to the quality of the machine translation, a more suitable metric for the quality assessment of the speech translation output could be designed. The fluency-adequacy metrics used in this research proved to be imperfect in terms of categories like punctuation and capitalisation, which do not directly affect the quality of the speech translation output itself. Also, sometimes it is not clear into which category an error should be classified, so the system could be simplified or made more transparent. More consistency can be achieved with proper instructions for error annotation and proper training of assessors. The T2S technology should also be analysed and assessed based on the naturalness, prosody and sound quality of the final speech production. Due to the limited scope of the present research itself, the ASR and T2S technology could not be discussed here in further detail. Also, the interface design, accessibility, user experience and ease of use should also be taken into consideration when assessing the overall app performance. All of these factors should be integrated in a comprehensive model of S2S translation and usability assessment.

This research has shown that ILA performs relatively well, but there is still plenty of room for S2S translation solutions to improve in order to allow for a smooth interaction between MT and humans. To meet the challenges of our technologically advanced times and the rapidly growing demand, "interfaces for

speech translation must balance competing goals: we want maximum speed and transparency (minimum interference) on one hand, while maintaining maximum accuracy and naturalness on the other" (Seligman and Waibel 2019: 221). In order to achieve those goals, adequate TQA metrics for S2S translation need to be developed to measure the overall quality for all stages of the process and set the desired benchmarks and thresholds for an end-to-end assessment of speech-to-speech translation and its usability in MT-mediated communication.

References

Arora, Karunesh, Arora, Sunita and Roy, Mukund Kumar. 2013. "Speech to Speech Translation: A Communication Boon". *CSI Transactions on ICT* 1(3): 207-213. https://doi.org/10.1007/s40012-013-0014-4.

Banchs, Rafael E., D'Haro, Luis F. and Li, Haizhou. 2015. "Adequacy-Fluency Metrics: evaluating MT in the continuous space model framework". *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23(3): 472-482. https://doi.org/10.1109/TASLP.2015.2405751.

Bawden, Rachel, Bilinski, Eric, Lavergne, Thomas and Rosset, Sophie. 2021. "DiaBLa: a corpus of bilingual spontaneous written dialogues for machine translation". *Language Resources and Evaluation* 55: 635–660. https://doi.org/10.1007/s10579-020-09514-4.

Callison-Burch, Miles Osborne, Chris and Koehn, Phillip. 2006. "Re-evaluating the Role of BLEU in Machine Translation Research". *11th Conference of the European Chapter of the Association for Computational Linguistics: EACL 2006*. 249–256. https://aclanthology.org/E06-1032.pdf.

Castilho, Sheila, Doherty, Stephen, Gaspari, Federico and Moorkens, Joss. 2018. "Approaches to Human and Machine Translation Quality Assessment: From Principles to Practice". In: Moorkens, Joss, Castilho, Sheila, Gaspari, Federico and Doherty, Stephen (eds.), *Translation Quality Assessment: From Principles to Practice.* Cham: Springer International Publishing. 9-38.

Chen, Xuewei, Acosta, Sandra and Barry, Adam E. 2017. "Machine or Human? Evaluating the Quality of a Language Translation Mobile App for Diabetes Education Material". *JMIR Diabetes* 2(1): e13. https://doi.org/10.2196/diabetes.7446.

Daems, Joke, Lieve Macken. 2013. "Annotation Guidelines for English-Dutch Translation Quality Assessment". Version 1.0. LT3Technical Report - LT3 13.02.

http://users.ugent.be/~jvdaems/TQA_guidelines_2.0.html. Accessed on: 9 November 2021.

Daems, Joke, Macken, Lieve and Vandepitte, Sonia. 2014. "Two sides of the same coin: assessing translation quality in two steps through adequacy and acceptability error analysis". *LREC 2014 Ninth International Conference on Language Resources and Evaluation.* 63-71.

Doherty, Stephen. 2016. "The impact of translation technologies on the process and product of translation". *International Journal of Communication* 10: 947-969.

Federmann, Christian and Lewis, William D. 2016. "Microsoft speech language translation (MSLT) corpus: The IWSLT 2016 release for English, French and German*". Proceedings of IWLST 2016.*

Hamon, Olivier, Mostefa, Djamel and Choukri, Khalid .2007. "End-to-End Evaluation of a Speech-to-Speech Translation System in TC-STAR". *Proceedings of Machine Translation Summit XI: Papers*. https://aclanthology.org/2007.mtsummit-papers.30.pdf. Accessed on: 21 June 2021.

Koehn, Philipp and Monz, Christof. 2006. "Manual and Automatic Evaluation of Machine Translation between European Languages". *Proceedings of the Workshop on Statistical Machine Translation*. 102-121. https://aclanthology.org/W06-3114.pdf. Accessed on: 21 June 2021.

Koehn, Philipp .2009. *Statistical Machine Translation.* Cambridge: Cambridge University Press.

Le, Ngoc-Tien, Lecouteux, Benjamin and Besacier, Laurent. 2018. "Automatic Quality Assessment for Speech Translation Using Joint ASR and MT Features". *Machine Translation* 32: 325-351. https://doi.org/10.1007/s10590-018-9218-6.

Lekić, Martina. 2021. *Vrednovanje odabranih aplikacija za prevođenje govora za jezične kombinacije en-de-hr*. Unpublished MA Thesis. Osijek: University of Osijek.

Linguistic Data Consortium. 2005. *Linguistic Data Annotation Specification Assessment of Fluency and Adequacy in Translations*. Revision 1.5.

Moorkens, Joos, Castilho, Sheila, Gaspari, Federico and Doherty, Stephen (eds.). 2018. *Translation Quality Assessment: From Principles to Practice.* Cham: Springer International Publishing.

Müller, Markus, Nguyen, Thai Son, Niehues, Jan, Cho, Eunah, Kruger, Bastian, Ha, Thanh-Le, Kilgour, Kevin, Sperber, Matthias, Mediani, Mohammed, Stüker, Sebastian and Waibel, Alex. 2016a. "Lecture translator speech translation framework for simultaneous

lecture translation". *Proceedings of NAACL-HLT of the Association for Computational Linguistics.* 82-86.

Müller, Markus, Fünfer, Sarah, Stüker, Sebastian and Waibel, Alex. 2016b. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).*

Papineni, Kishore, Roukos, Salim, Ward, Todd and Zhu, Wei-Jing. 2002. "BLEU: a Method for Automatic Evaluation of Machine Translation". *Proceedings of the 40th Annual Meeting of the ACL* 40. 311-318.

Seligman, Mark and Waibel, Alex. 2019. "Advances in Speech-to-Speech Translation Technologies". In: Ji, Meng and Oakes, Michael (eds.), *Advances in Empirical Translation Studies: Developing Translation Resources and Technologies*. Cambridge: Cambridge University Press. 217-251.

Sperber, Matthias and Paulik, Matthias. 2020. "Speech translation End-to-End promise: taking stock of where we are". *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7409-7421. https://doi.org/10.18653/v1/2020.acl-main.661.

Takezawa, Toshiyuki, Kikui, Genichiro, Mizushima, Masahide and Sumita, Eiichiro. 2007. "Multilingual spoken language corpus development for communication research". *Computational Linguistics and Chinese Language Processing* 12(3): 303–324.

TAUS. 2016. "TAUS Post-Editing Guidelines". https://www.taus.net/think-tank/articles/postedit-articles/taus-post-editing-guidelines. Accessed on: 21 June 2021.

Waibel, Alex and Fügen, Christian. 2008. "Spoken language translation. Enabling cross-lingual human-human communication". *IEEE Signal Processing Magazine* 25: 70-79.

Wahlster, Wolfgang. 2000. "Mobile speech-to-speech translation of spontaneous dialogs: An overview of the final Verbmobil system". In: Wahlster, Wolfgang (ed.), *Verbmobil: Foundations of Speech-to-Speech Translation.* Berlin, Heidelberg: Springer. 3-21.

PROCJENA KVALITETE STROJNOG PRIJEVODA GOVORA: STUDIJA SLUČAJA APLIKACIJE ILA

Sažetak

Strojno je prevođenje sve kvalitetnije i sve je više prisutno u svakodnevnom životu. Zbog porasta potražnje za brzim i pristupačnim prijevodima teksta i govora, strojno se prevođenje nameće kao općeprihvaćeno rješenje, što dovodi do korjenitih

promjena i prilagodbi u prevoditeljskoj struci i praksi te istodobno višejezičnu komunikaciju čini lakšom nego ikada do sada. Ovaj se rad bavi aplikacijom Instant Language Assistant (ILA) za strojni prijevod govora. ILA omogućuje višejezičnu komunikaciju posredovanu strojnim prevođenjem, a temelji se na najnovijim tehnološkim dostignućima, i to na automatskom prepoznavanju govora, strojnom prevođenju i sintezi teksta u govor. Cilj je rada procijeniti kvalitetu prijevoda razgovornog jezika dobivenog pomoću aplikacije ILA i to za parove jezika engleski – njemački te engleski – hrvatski. Kvaliteta prijevoda analizira se u nekoliko faza: kvalitetu prijevoda procjenjuju stručnjaci pomoću metode procjene tečnosti i točnosti (engl. Fluency/Adequacy Metrics*), zatim se provodi ograničena redaktura strojno prevedenih govora (engl.* light post-editing*), nakon čega slijedi automatsko vrednovanje strojnog prijevoda (*BLEU*). Strojno prevedeni govor procjenjuje se i uzevši u obzir o kojem je jezičnom paru riječ kako bi se dobio uvid u to utječu li jezični parovi na strojni prijevod i na koji način. Rezultati pokazuju da su prijevodi dobiveni pomoću aplikacije ILA za strojni prijevod govora procijenjeni kao razmjerno visokokvalitetni bez obzira na metodu procjene, kao i da se ljudske procjene kvalitete prijevoda poklapaju sa strojnima.*

Ključne riječi: tehnologija prijevoda govora, aplikacije za strojni prijevod govora, procjena kvalitete prijevoda, aplikacija ILA