Assessment of Selected Speech Translation Apps for en-de-hr Language Pairs

Lekić, Martina

Sveučilište J. J. Strossmayera u Osijeku

Filozofski fakultet

Diplomski studij engleskog jezika – prevoditeljski smjer i njemačkog jezika – prevoditeljski smjer

Martina Lekić

Vrednovanje odabranih aplikacija za prevođenje govora za jezične kombinacije en-de-hr

Diplomski rad

Mentorica: prof. dr. sc. Marija Omazić

Osijek, 2021.

Sveučilište J. J. Strossmayera u Osijeku

Filozofski fakultet

Diplomski studij engleskog jezika – prevoditeljski smjer i njemačkog jezika – prevoditeljski smjer

Martina Lekić

Vrednovanje odabranih aplikacija za prevođenje govora za jezične kombinacije en-de-hr

Diplomski rad

Znanstveno područje: humanističke znanosti

Znanstveno polje: filologija

Znanstvena grana: anglistika

Mentorica: prof. dr. sc. Marija Omazić

Osijek, 2021.

J.J. Strossmayer University of Osijek

Faculty of Humanities and Social Sciences

Double Major MA Study Programme in English Language and Literature – English Translation and Interpreting Studies and German Language and Literature – German Translation and Interpreting Studies

Martina Lekić

Assessment of Selected Speech Translation Apps for en-de-hr Language Pairs

Master's Thesis

Supervisor: Dr. Marija Omazić, Professor of Linguistics

Osijek, 2021

J.J. Strossmayer University of Osijek

Faculty of Humanities and Social Sciences

Double Major MA Study Programme in English Language and Literature – English Translation and Interpreting Studies and German Language and Literature – German Translation and Interpreting Studies

Martina Lekić

Assessment of Selected Speech Translation Apps for en-de-hr Language Pairs

Master's Thesis

Scientific area: humanities

Scientific field: philology

Scientific branch: English studies

Supervisor: Dr. Marija Omazić, Professor of Linguistics

Osijek, 2021

IZJAVA

Izjavljujem s punom materijalnom i moralnom odgovornošću da sam ovaj rad samostalno napravila te da u njemu nema kopiranih ili prepisanih dijelova teksta tuđih radova, a da nisu označeni kao citati s napisanim izvorom odakle su preneseni. Svojim vlastoručnim potpisom potvrđujem da sam suglasna da Filozofski fakultet Osijek trajno pohrani i javno objavi ovaj moj rad u internetskoj bazi završnih i diplomskih radova knjižnice Filozofskog fakulteta Osijek, knjižnice Sveučilišta Josipa Jurja Strossmayera u Osijeku i Nacionalne i sveučilišne knjižnice u Zagrebu.

U Osijeku, 2021.

Martina Lekić

Martina Lekić, 0122224005

ACKNOWLEDGEMENTS

I would first like to thank my supervisor, Dr. Marija Omazić, whose valuable guidance directed me throughout my studies and whose passion has driven me to write this master's thesis and enjoy the process.

I would like to acknowledge TranslateLive CEO Peter Hayes, who provided us with a free access to the ILA app, making this whole research possible. Thank you for your willingness to even further collaborate with the University of Osijek.

I would like to thank my professors Dubravka Vidaković Erdeljić and Marija Viljušić, as well as my colleagues Talia Jurić, Anamarija Pacek, Anja Jokić and Hrvoje Vlainić for their participation and help in this research.

I would like to thank my parents, brothers and my sister Ivana for the irreplaceable care and support during the last five years.

I would like to thank my husband Matija, whose invaluable encouragement, understanding and love provided me with everything I needed to become the person I am today.

I offer deep and humble gratitude to God, without whom nothing of this would be possible!

Contents

1. Introduction

In a constantly developing and globalizing world, technological advances have brought a major shift in translation as a means of cross-lingual communication. Computer-assisted translation (CAT) tools and machine translation "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). With people around the world coming together into one big global family, the need for a lingua franca becomes indispensable. Even though a lingua franca that would bring the whole world together under one universal language does not exist, translation is the age-old solution to the age-old problem of the language barrier. It is said that since there is humans, there is language, and since there is language, there is translation. Enabling communication between people who speak different languages, translation plays a tremendous role in human history. And today, translation only gains increased importance. Waibel and Fügen argue that "the expanding interest and excitement can be explained by a convergence of emerging and powerful new technical capabilities and a growing appreciation of the needs for better cross-lingual communication in a globalizing world" (2008:70). In other words, technological advancements enabling disruptive translation solutions have finally appeared.

One of the futuristic aspects of translation, which slowly enters into every domain of human lives, is automated speech translation. Having a machine translating text from one language into another in a matter of seconds is an already widely accepted feature. However, speech translation, or to go even a step further, Speech-to-Speech (S2S) translation involves some impressive state-of-the-art technology in order to produce spoken output in the target language from spoken input in the source language.

This paper aims to conduct research of one of the S2S translation apps – the Instant Language Assistant. Bringing together all of the state-of-the-art translation technology, namely, Automatic Speech Recognition (ASR), Machine Translation (MT) and Text-to-Speech Synthesis (TSS), ILA functions as a language mediator between people regardless of the language they speak.

1.1. Rationale

Living in an instant world requires instant solutions to everyday problems. MT has proved to be an excellent tool in helping not only professionals in their work but also laypersons, enabling them to function on an everyday basis in the cross-lingual global society. Having in mind the great

challenges posed to the translation industry by an ever-growing demand for quick, cheap and accurate translations, the goal of this paper is to look into the potential of speech translation apps to eliminate the need for human interpreters in everyday situations.

1.2. Research Questions

With the boom in translation technologies and new solutions being introduced every day, one decisive question is being raised: *Are translation technologies to be trusted?* Thinking about the advantages of producing translations without boundaries such as physical needs and limited capacities of human translators, MT gains more and more trust. On the other hand, the ability of a machine to "think" and produce spot on translations without human intervention is still doubted. This paper aims to answer a more specific question, namely *What is the quality of translations of conversational language produced by S2S translation apps?* Another issue with MT is the disproportion in the quality of the translated output depending on the language pair. So this paper will try to answer the question: *On what levels does the quality of the ILA-produced translations vary based on the language pair?* Similarly, the translation quality often depends on the type and topic of the source text. This leads to the final question: *Is the quality of the produced translation dependent on the situation?*

1.3. Research Design

Since the idea behind the creation of ILA was to provide an instant language assistant for everyday real-life situations, this research will be performed on three dialogues comprised specifically for this purpose. Each dialogue takes place in a different real-life setting – in the bank, at the hotel and in the grocery store, and each of the dialogues is carried out in two language combinations – English and German; as well as English and Croatian.

In the second chapter of the paper, speech translation in general is introduced, giving a brief overview of the development of speech technologies, the benefits and challenges of speech translation apps, as well as a look into the future of speech translation. Moving on to the third chapter, the S2S translation app ILA is introduced. This includes a clarification of what ILA is, how it works and a discussion on the benefits of the ILA app. In the analytical part of the paper, the ILA was tested on several levels. The analysed material consists of the three real-life dialogues each originally comprised in the language combination English-German and English-Croatian. The dialogues were dictated to the ILA app and the results were documented for this research. Each translation was submitted to a detailed quality assessment based on two levels – Fluency and Adequacy, conducted by professors of translation at the Faculty of Humanities and Social Sciences

in Osijek. Then, the ILA-produced translations were lightly post-edited, followed by a discussion of post-editing results. In addition, the ILA-produced translations were assessed by the Automated Translation Metrics, giving a fully objective, automatic quality assessment of the MT. The MTs were then compared to human translations, produced by fellow students, graduates of the MA in Translation Studies in Osijek using a modified translation-marking grid. Lastly, all of the results were compared from two perspectives – firstly, with regards to the language combination they were performed in; and secondly, with regards to the situation in which the dialogues were performed.

2. Speech Translation

Speech translation (ST) as a concept seemed to be a purely futuristic idea just a few decades ago, as Seligman et al. (2017) suggests, not far behind the videophone and the flying car, both of which are already here. The early 1970's science fiction franchise Star Trek gave an idea of how this may look like in the year of 2260, introducing the Universal Translator, a revolutionary device enabling the crew of the ship to communicate with any extra-terrestrial species in the universe. Back then, it was hardly imaginable that in the early 2000, we will already have devices, which perform similar actions. Another early depiction of the speech translator was displayed in Douglas Adams' *Hitchhiker's Guide to the Galaxy* from 1979. Namely, the Babel Fish, the universal translator, which simultaneously translates from one spoken language into another directly into the host's brain wave matrix, was introduced. Even though it seemed far away, speech technologies have already spread widely into every sphere of life, becoming one of the technological solutions that shape our present and will surely shape our future. Speech translation brings together various components of our technologically advanced times. "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 and machine translation" (Sperber and Paulik 2020:7409). In comparison to Speech-to-Text translation, Speech-to-Speech translation adds another important component to generate the translated text into target language speech and that is Text-to-Speech Synthesis (TSS). Before speech translation in general could be achieved, some basic speech technologies had to be developed.

2.1. Development of Speech Technologies

Milestones in the development of speech technologies date far back into history. Stein (2013) argues that probably the first thoughts on MT emerged out of two philosophical schools dealing with the nature of language. One of them was focused on creating secret encoded languages and the other evolved around the idea of a universal language that would allow communication without barriers (Stein 2013). In the 17th century, a German philosopher and mathematician Gottfried Wilhelm Leibnitz built on the idea of a universal language in his theory of monades as he tried to develop a set of "termini primi" or smallest units of meaning to compose all thinkable thoughts (Stein 2013). In 1773, Christian Kratzenstein, a Russian scientist and psychology professor built a device which produced sounds similar to human vowels using resonance tubes connected to organ pipes (Moskvitch 2017). Over a decade later in Vienna, Wolfgang von Kempelen created the Acoustic-Mechanical Speech Machine, a model of the human

vocal tract that imitates the process of speech synthesis. In the early 19th century, Kempelen's system was improved by the English inventor Charles Wheatstone. Another huge step forward in speech recording was Thomas Edison's "Dictaphone", patented in 1907 (Moskvitch 2017). A few decades later, in World War II, the decipherment of the German Enigma code in Bletchley Park laid the foundations for practical MT (Stein 2013). Regarding this experience, Warren Weaver wrote: "[…] it is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the 'Chinese Code'. If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation? (Weaver, as cited in Stein, 2013:6)". This is considered the birth of MT.

In 1953, the first Automatic Digit Recognition machine came along. Audrey could recognise spoken digits – zero to nine – with more than 90% accuracy, but only when spoken by its inventor and with considerably less accuracy when spoken by unfamiliar voices (Moskvitch 2017). In 1962, IBM introduced the Shoebox, a machine that could understand 16 English words. Until 1996, large amounts of money were spent in order to develop MT systems (Stein 2013).

In 1971, the US Department Defence's research agency DARPA funded the Speech Understanding Research programme (Moskvitch 2017). Leading companies like IBM and academia such as Carnegie Mellon University (CMU) and Stanford Research Institute joined their forces and Harpy was born. Harpy could recognise 1,011 words as well as whole sentences, which helped reduce speech recognition errors. As speech recognition systems evolved further, IBM introduced the voice-activated typewriter Tangora with a 20,000-word vocabulary (Moskvitch 2017). With their own approach and technological advances, IBM's competitors Dragon Systems released Dragon NaturallySpeaking in 1997. Unlike its predecessors, it was the first continuous speech recognition product making machines more human-like. From the beginning of the 1980s and with the development of speech technologies, MT experienced continuously increasing popularity (Stein 2013). In 1983, the Japanese company NEC demonstrated the earliest speech-translation system. Limited to domain-restricted phrasebooks, it illustrated the vision of automatic speech interpretation (Seligman et al. 2019). With the development of the main components of any speech translation system such as speech recognition, MT and speech synthesis, speech translation gained pace towards the state-of-the-art advancements in S2S translation.

In 1992, ATR, Carnegie Mellon University (CMU), the Karlsruhe Institute of Technology (KIT) and Siemens established the Consortium for Speech Translation Advanced Research, or the C-STAR. In 1993, the group demonstrated the first SLT, showing voice-to-voice rendering (Seligman and Waibel 2019). In 1998, the first practical demonstration of unrestricted or open-ended speech translation took place. This resulted in an SLT system, which enables interactive

correction of ASR errors, leading SLT one-step further. The same year, Germany had its speech translation product for PCs. *Talk & translate* worked for German and English, the translations were added Via Voice from IBM and its associated text-to-speech system and it required a twenty-minute training session before usage (Seligman and Waibel 2019). Another great project in Germany during the 1990s was the Verbmobil project, which further developed statistical machine translation (SMT) to create the first statistical speech translator (Seligman and Waibel 2019). Given its learning ability and consistency, SMT often resulted in better translation quality than rule-based methods. In 2004, DARPA launched several research programs to develop speech translation for governmental use (Seligman and Waibel 2019). This resulted in "two-way" speech translators. In 2006, first two spoken language translation (SLT) products for telephony entered the Japanese market working in the language combination Japanese and English (Seligman and Waibel 2019).

In the last decade, "machine learning techniques loosely based on the workings of the human brain have allowed computers to be trained on huge datasets of speech, enabling excellent recognition across many people using many different accents" (Moskvitch 2017). As big data grew ever bigger and the market rapidly expanded, Google Translate took off in 2006-2007. In that respect, free text translation became available to every internet user, while rule-based translation systems switched to statistical MT (Seligman et al. 2017). In addition, with the advent of mobile phones to smartphones and the mobile app market forming, advanced speech and MT technology could now fit on a phone (Seligman et al. 2017). In 2009, Mobile Technologies launched Jibbigo, "the first speech translator to run without network assistance on iPhone and Android smartphones" (Eck et al., as cited in Seligman at al. 2017:14). Jibbigo featured a 40,000-word vocabulary and produced voice output from voice input faster than a message could be typed. The first app provided speech translation in the language combination English and Spanish, but added fifteen languages in the following couple of years. In 2013, the company was acquired by Facebook (Seligman et al. 2017). In 2010, Google entered the SLT field with network-based mobile speech translation, demonstrating the Conversation Mode. It started with English and Spanish, but expanded to fourteen languages in a year. Four years later, Microsoft launched the Skype Translator "judging that exploitation of neural networks for ASR and MT had finally brought S2ST past the usability threshold" (Seligman and Waibel 2019:228). In the 2010s, the S-MINDS system was released. It was used in healthcare and relied on pre-translated phrases.

Another great boom in the development of speech technologies was the Google Voice Search app for the iPhone. Quickly, Apple offered their version, Siri; Microsoft called its AI Cortana and Amazon introduced Alexa (Moskvitch 2017). In 2017, Google introduced the

Translatotron, an End-to-End, Speech-to-Speech translation model that retains the voice of the original speaker after the translation (Jia and Weiss 2019), making the MT more and more transparent. In 2019, Google introduced interpreter mode for its home devices. Saying: "Hey, Google, be my French interpreter", will activate spoken and text translation on smartphones. (Kohn 2019). One of the state-of-the-art speech translation systems widely used today is Microsoft Presentation Translator, running a real time transcription of the speaker's words and broadcasting the translation into multiple languages. Another one is the first simultaneous interpreting service InterACT operating in lecture halls of KIT (Seligman et. al 2017). Of high importance are various European Union pilots supporting human interpreters by automatically generating terminology and listing numbers and names, which are difficult to remember while interpreting.

Although the technology necessary for speech translation has finally emerged from science fiction, research and forecasts, despite its usability potential, many remain sceptical (Seligman and Waibel 2019).

2.2. Benefits of Speech Translation Apps

In a constantly growing and globalizing community, communication is one of the basic principles that brings together people, cultures, ideas, knowledge and many more. Instead of searching for ways to make all people come together under a common lingua franca, speech translation apps function as a language mediator between people regardless the language barrier. In this way, people preserve their cultural background and features that differentiate them from others but are able to communicate with anyone just as with a fellow citizen.

Speech translation apps are instant simultaneous interpreters, accompanying people on all of their journeys, making them feel safe and resourceful regardless the circumstance. Just like having a simultaneous interpreter with the knowledge of any language wherever you go, but much simpler and cheaper. ST apps bridge the gap between dialogue participants with minimal interference and latency. They transcribe the spoken input as it has been said, and translate it just a few moments later. Many of them have the option of reading the target language translation aloud. Moreover, speech translation apps are independent and unbiased, guaranteeing objectivity in every situation. The user can rely on his app for a faithful display of what has been said. Most importantly, speech translation apps are available to everyone, removing all obstacles in way of easy communication. The only precondition is to have a smartphone (smartwatch, smart glasses), and internet access, but in our day and age, that is almost no problem.

With the possibility to write down the input message instead of speaking it, as well as reading the output translation instead of listening to it, speech translation apps are suitable for the hearing or seeing impaired, giving them the possibility to communicate just as everybody else. As Doherty (2016) concludes: "These technologies 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 2016:1).

2.3. Challenges of Speech Translation Apps

Due to its multi-faceted design and workflow, speech translation apps are prone to various difficulties in the production of translation. Clearly, „better ASR, MT, or TTS performance makes for better speech translation performance" (Waibel and Fügen 2008:70). Functioning on three separate levels, a minor mistake in the first step of the translation process can become a quite serious one by the end of the translation process. The challenges are divided into three subsections according to each task that a speech translation app performs. The first task is automatic speech recognition where speech is recognized and transcribed. In the second step, machine translation is performed, translating the source language text into text in the target language. Lastly, text-to-speech synthesis creates the speech signal from text (Arora et all. 2013:209).

2.3.1. Automatic Speech Recognition (ASR)

The fact that we speak considerably different than we write or read, not to mention the difference between how we sound and how we hear ourselves, is one of the main challenges for any automatic speech recognition system. From the way we pronounce words to pace, every feature of the speaker's way of speaking influences the speech recognition and at the same time, the final translation product.

Some distinctive traits of free spontaneous human speech are false starts, hesitations, repetitions, spontaneous speech and disfluency (Waibel and Fügen 2008). All of these easily confuse the system, which will therefore produce inaccurate results or no result at all. The above-mentioned pronunciation is a feature shaped by not only our origin or background, but the physiognomy of our vocal organs as well. The first will determine how close to the standardized language the speaker speaks, whether he has a strong or a weak accent, whether he has a regional variation or if he uses cross-language expressions. The latter predefines how clearly a speaker articulates voices that is, how well an ASR system will be able to recognize the speech production.

Speaking style can also differ depending on the speech type. For example, the speech of a TV anchor is mostly read without hesitations and disfluencies and can therefore be recognized with high accuracy (Seligman et al. 2017). The more informal it gets, the harder it is for an ASR system to recognize accurately what is said since the speech becomes more casual and spontaneous. Lectures are for example pre-prepared speeches, but the speech flow can be disrupted by digressions, questions and clarifications, making it difficult for the machine to recognize the content correctly. Dialogues are unprepared, spontaneous and free conversations and the dynamic can be affected by the topic, agreement or disagreement between parties and so on, making it difficult for the machine to keep up. Another feature closely connected to this one is pace. The same principle applied in consecutive interpreting should be applied here as well. As opposed to simultaneous interpreting, where listening and translating are performed in parallel while the speaker keeps speaking, in consecutive speech interpretation, "a speaker pauses after speaking to give the system (or the human interpreter) a chance to produce the translation" (Seligman et al. 2017:46).

Another thing that greatly affects accuracy of speech recognition is disfluency in speech production. When the speaker is unprepared or cannot produce fluent speech, the machine is easily confused. Speed and latency also pose potential difficulties to the speech recognition system. When speaking too fast, the speaker tends to slur, shorten or merge words, making it much harder for the ARS to recognize correctly what is said. On the other hand, excessive latency, that is waiting time, may be challenging especially if it appears in the middle of long, syntactically difficult sentences where, for example, the verb occurs at the end of the sentence rather than at the beginning. Range is another feature with a great impact on the ASR system's productivity. Free conversations are harder to follow due to the wide range of occurring vocabulary, syntactic structures, grammar, common sense, implications etc. In restricted-domain dialogues, the system is much more likely to accurately recognize an ambiguous word based on the domain in which it is used.

Moreover, an ASR system cannot understand the means of illocution, intonation and so on, simplifying the input to declarative statements, with no means to distinguish the speaker's intention. Furthermore, some external factors may influence the accuracy of speech recognition, such as environmental noise or issues deriving from microphone positioning and usage (Waibel and Fügen 2008).

2.3.2. Machine Translation (MT)

Probably the greatest stumbling block in the way of perfect machine translations is the inability of a machine to understand context. Similarly, Müller concludes, "Despite a difference in the overall quality of the translations, MT systems suffer from not being able to anticipate context like human interpreters" (Müller et al. 2016:83). This can be especially confusing in non-restricted domain dialogues because the interlocutors often shift from one topic to another, making it impossible for the MT to realize and act, that is, translate accordingly. Being trained on a specific set of sentences or translation memories, machine translation cannot move outside those boundaries, use common sense to connect certain ideas or understand the speaker's intention.

Another challenge when it comes to machine translation is its dependency on human intervention. "Despite the widespread and diverse adoption of MT in research and practice, most machine translated content still requires some form of human intervention to edit the MT output to the desired level of quality and/or to verify its quality before publication, dissemination, product release, legal compliance, and so on" (Doherty 2016:958). In other words, even though machine translation adds precious speed to the translation industry, one still cannot completely rely on the machine itself. "MT, however, is not without its own risks to quality, misrepresentation, and misuse, and it presents another force that translators must contend with as the fixing of machine-translated output becomes the bread and butter of many professional translators" (Doherty 2016:962).

2.3.3. Text-to-Speech Synthesis (TSS)

When it comes to text-to-speech synthesis, one of the difficulties, which comes to mind is the discrepancy between written and spoken language. The question is also, how skilfully a machine can reconcile these differences. Developing on the idea that machine translation cannot understand context, systems for speech synthesis cannot adapt the translation to the register, range or speaking conventions. Therefore, the target language text is often general and nonspecific.

Speech translation technologies offer a wide range of improvements in the translation industry. Although improvements in the quality of speech-to-speech translation technologies are evident, even the best S2ST systems require some degree of human intervention (Doherty 2016). One of the greatest challenges for speech translation apps in the future will be managing to correct human mistakes and produce impeccable translations regardless of the given material. Waibel and Fügen conclude, "To err is human, and useful systems must accommodate a speaker's mistakes" (Waibel and Fügen 2008:71).

2.4. The Future of Speech Translation

Seligman (2017) describes the goal of S2ST technologies as "maximum speed and transparency (minimum interference) on one hand, while maintaining maximum accuracy and naturalness on the other" (Seligman et al. 2017:9). Much research, ideas and visions seem to be necessary before this could be achieved. However, some futuristic aspects of speech translation do not seem so far away. One obvious and already present trend is the technology's migration to mobile and convenient platforms (Seligman et al. 2017). Many free translation apps are already available for iPhone and Android. For example, Google Translator synthesizes the translation into speech although it sounds quite robotic. Nevertheless, Microsoft Translator app synthesizes the translation into natural sounding speech, even for not so widely spread languages like Croatian. Meanwhile, translating wristbands and smart watches are increasing in popularity, and some of the watches can even exchange translations with a nearby smartphone.

Moving further from the wrist, a prototype of an eyeglass-based delivery of real-time translation has been introduced (Seligman et al. 2017). In the near future, smart glasses will offer not only speech translation, but will be able to instantly translate signs and other written material as well. Alongside wristbands and glasses, some companies offer translating necklaces and earpieces with a built-in translation software. Nevertheless, it certainly will not stop there. Soon, speech translation will be embedded into everyday objects, just as Google's interpret mode for home devices.

3. ILA – The Instant Language Assistant

Speaking multiple languages is not a skill all people have. Even if they do, at some point in their lives they will come across someone with whom they do not share a common language. Offering a glimpse into the future of translation devices, TranslateLive has created an innovative software and hardware solution with the capability to connect everyone regardless of the language barrier[1]. ILA or the Instant Language Assistant is a S2ST app designed to perform the role of a language mediator between people who speak different languages but need to communicate to each other. The director of development for TranslateLive, Jim Holmes, describes ILA as "the product of the future for translation, travel and global human interaction"[2].

Image 1: The ILA device

3.1. About ILA

The initial idea of the TranslateLive App was introduced in February 2017. After years of hard work, in March 2019, the first ILA prototype was launched. Rising to meet the challenges in an age of instant gratification, ILA devices seek to enhance the global communication experience[3]. For this purpose, TranslateLive has created two devices specialized for specific usage situations: The ILA Traveller and the ILA Pro device. ILA Traveller is small and compact, so it can accompany users on their journeys. When in need of translation, the user pulls the device out and

[1] https://www.translatelive.com/ila-solutions/
[2] https://www.indiegogo.com/projects/ila-the-instant-language-assistant-device#/
[3] https://console.cloud.google.com/marketplace/details/google-cloud-platform/compute-engine?pli=1

hands it to the other person. Whether it is ordering food or asking for directions, ILA allows its user to speak any language, enjoying the travel experience to the fullest. 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. The ILA adds real-time communication capabilities to any business[4]. With more than 120 languages and dialects, the Instant Language Assistant assists its users anywhere and in any situation.

[4] https://www.translatelive.com

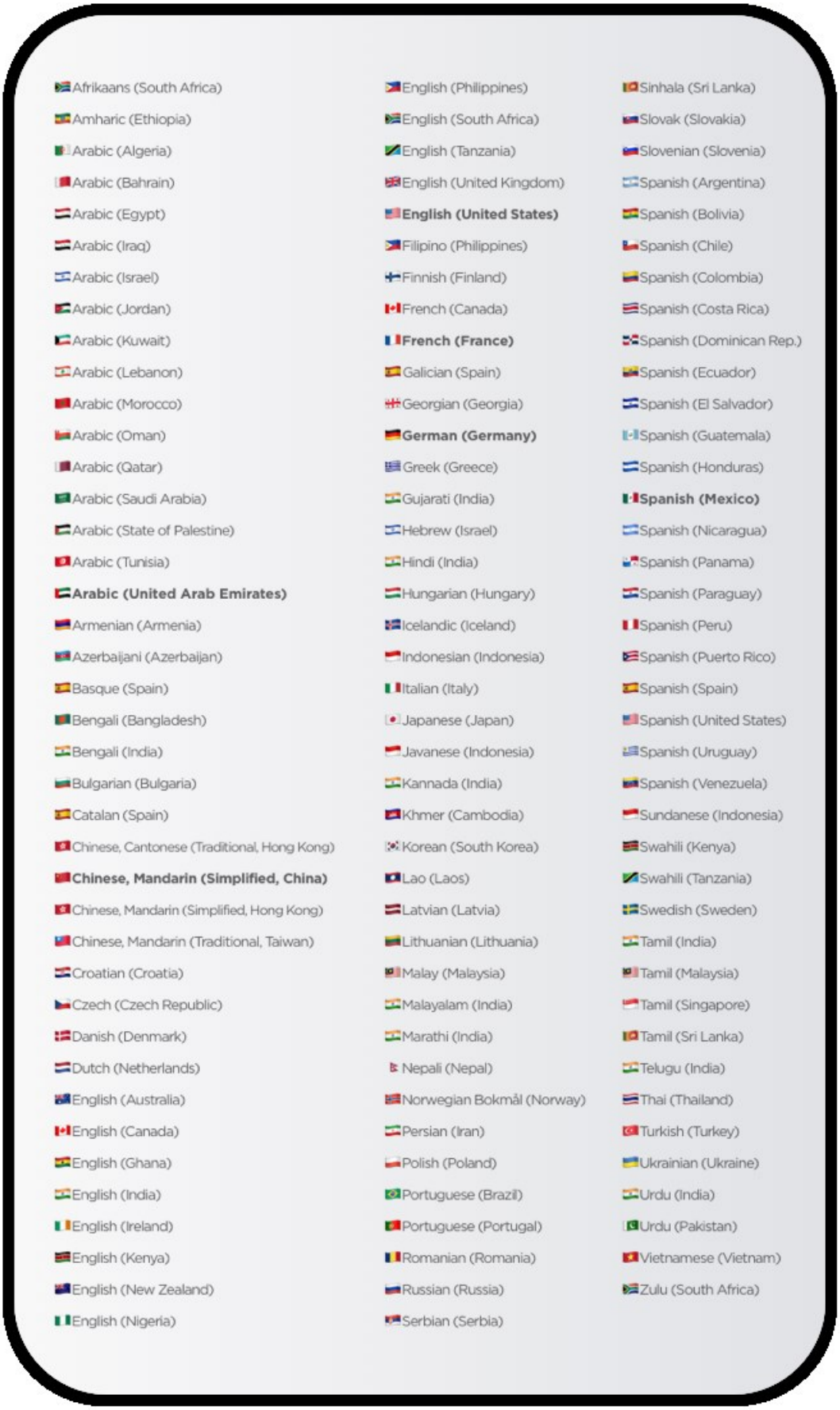

Image 2: Languages included into the ILA app

3.2. How Does it Work?

In order to start translating, all a user has to do is press the button and start speaking in her/his preferred language. The speech will show up in text format, so the speaker can ensure everything is written correctly. The translated text will appear to the person on the other side in their selected language, with the option to be read out aloud. Then, the second person presses the button to respond.

The TranslateLive platform 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"[5], 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 accurately converts speech into text format using an API powered by Google's AI technologies[6]. Live automated language translation uses state-of-the-art TranslateLive artificial intelligence[7]. The AI part of the solution uses engines as Google, Microsoft, Amazon and Apptek for non-specifically trained customer systems.

3.3. Benefits of the ILA

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 the 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. Providing both the possibility to speak or write what you want to say as well as to read and hear the translation, makes it possible for them to successfully communicate with anyone as well. Another advantageous feature is privacy, since all the conversations are encrypted, private and HIPAA compliant. Last but not the least, the fact that each speaker is enabled to check the speech on their own screen, ensures a much higher accuracy rate, which is important for a successful communicative act[8].

[5] https://console.cloud.google.com/marketplace/details/google-cloud-platform/compute-engine?pli=1
[6] https://cloud.google.com/speech-to-text/
[7] https://www.translatelive.com/ila-solutions/
[8] https://www.translatelive.com/ila-solutions/

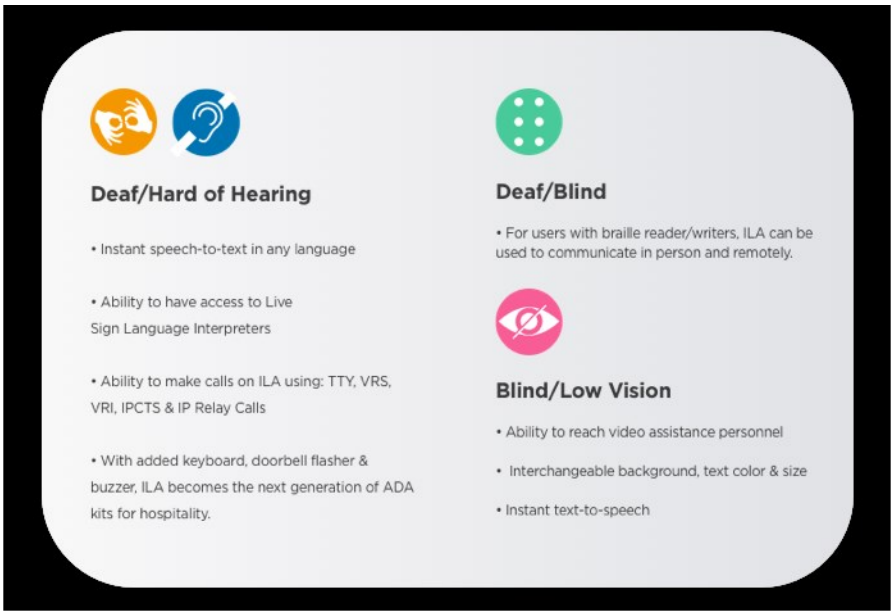

Image 3: Advantages of the ILA for people with disabilities

Alongside all of this, TranslateLive ensures the possibility to contact call centre help agents, available at all times, via the app, for users who speak a language not among the 120 languages provided by ILA. Also, there is an in-app broadcast mode, which allows public speakers to have their speech translated in different languages simultaneously, for listeners who want to follow along in their native language.

Another contribution of the ILA app to the community is related to the COVID-19 pandemic. Currently, the app is being trained based on specific strings in order to facilitate the vaccination process both for medical workers and patients who do not speak a commmon language. Questiones like *Is this your first dose of COVID-19 Vaccine?* and phrases like *Your arm may be sore for a day or two*, have been translated into six languages with many in process. Due to our collaboration in this project, those Covid strings are now translated into Croatian as well.

Undoubtedly, the Instant Language Assistant has numerous benefits for the individual, but for the global community as well. Peter Hayes, the TranslateLive CEO describes it in the following way: “We live, eat and breathe communication access. Our goal is to serve our community and breaking down barriers is what we have always done”. It is not only breaking down the language barriers between people, but it also brings people together into one big global family.

4. Testing of the ILA-produced Translations

The following chapter focuses on the testing of the translations produced via the ILA app. The translations were subjected to different procedures in order to gain a larger perspective about the overall quality of the Instant Language Assistant's output. Firstly, a quality assessment based on a metrics for human evaluation of MT was carried out. Another means of evaluating MT conducted on the translations is post-editing, followed by an automated translation metrics. The next step in evaluating ILA-produced translations is comparing the MT with HT of the dialogues using a specified marking grid. In the last part of the research, the results of all three evaluating methods were compared concerning the language pair and with the regards to the situation, that is the dialogue topic.

4.1. Quality Assessment of the ILA-produced Translations

Each of the three dialogues comprised for this research was evaluated independently both for the language pairs English-German and English-Croatian. This part of the analysis was conducted by three professors of translation studies at the Faculty of Humanities and Social Sciences in Osijek. All three professors are also practicing translators with 10-25 years of experience working with both translation and interpreting. The quality assessment is based on an Adequacy-Fluency Metrics for human evaluation of MT, shown in table 1 below.

4.1.1. Quality Assessment Methodology

The Adequacy-Fluency Metrics for evaluating MT is a two-dimensional evaluation metrics conducted by human translators or linguists, aiming at dissociating semantic and syntactic components of the translation process in order to provide a more balanced view on translation quality (Banchs et al. 2015). Even though it is a metrics for evaluating text-to-text translation, for lack of a more suitable metrics, in this research it is used to evaluate the translation output of the ILA app. In order to gain the most relevant results despite the shortage of a specialised evaluation metrics, the Adequacy-Fluency metrics had to be slightly adjusted to the needs of the assessment of the speech translation output. Evaluation categories like punctuation and capitalisation were therefore not taken into account, since they do not directly indicate a translation error, especially because of the ASR part of the speech-to-speech translation. However, in translations into German, capitalisation of nouns was considered as an indicator of the translation quality. When evaluating fluency, the evaluators have access only to the translation output. Assessing only the monolingual translation, the evaluators focus on the fluency of the produced target language translation. In other

words, the evaluators answer the question: *Is the language in the output fluent?* When evaluating adequacy, the evaluators have access to both the source and the target text. In this cross-language quality assessment, the evaluators are led by the question: *How much meaning is preserved?*, focusing on how faithfully the information given in the source text is translated into the target language.

Each evaluator was provided with the translation input, the MT output and a table with error categories for each of the evaluating levels. Starting with fluency, the evaluators marked and categorized each error found in the target text. They did the same when evaluating adequacy, marking all adequacy errors found in the comparison of the source and target text.

Table 1: Adequacy-Fluency Metrics

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	singular/plural	word non-existent	compound	repetition	logical problem
terminology	verb form		punctuation	disfluent	paragraph
quantity	article-noun agreement		typo	short sentences	inconsistency
time	noun-adjective agreement			long sentences	coherence – other
meaning shift caused by punctuation	subject-verb agreement			text type	
meaning shift caused by misplaced word	reference			style – other	
deletion	missing				
addition	word order				
explicitation	structure – other				
coherence	grammar – other				
inconsistent terminology					
other					

After categorizing the errors, the evaluators graded fluency and adequacy for each of the dialogue according to the grading scale given in Table 2.

Table 2: Fluency-Adequacy Metrics Grading Scale

Grade	Error number	Translation quality
excellent (5)	1-5	maximum/ publication standard
very good (4)	6-10	minimum professional standard
good (3)	11-15	adequate, standard
sufficient (2)	16-20	inadequate, substandard
insufficient (1)	21-25	completely inadequate

4.1.2. Quality Assessment Results

The results of the quality assessment are given separately for each of the two language pairs and for adequacy and fluency as well.

Grades given for the translation quality of dialogues led in English and German (Table 3 and 4) range from excellent (5) to good (3) both for fluency and for adequacy. The translation of the dialogue *At the hotel* received a maximum of three excellent (5) grades for adequacy. The translation of the dialogue *At the store* received the grade good (3) two times for fluency and once for adequacy. The average grade for translations of the dialogues *At the bank* and *At the hotel* is excellent (5) for fluency and for adequacy, indicating the maximum translation quality. The translation of the dialogue *At the store* received the grade good (3) for fluency and the grade very good (4) for adequacy denoting a minimum professional standard.

Table 3: Quality Assessment results for the language pair English and German

FLUENCY	**Evaluator 1**		**Evaluator 2**		**Evaluator 3**		
	Error No.	Grade	Error No.	Grade	Error No.	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**

Table 4: Quality Assessment results for the language pair English and German

ADEQUACY	**Evaluator 1**		**Evaluator 2**		**Evaluator 3**		
	Error No.	Grade	Error No.	Grade	Error No.	Grade	**Average grade**
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**

The grades given for the translation quality of dialogues led in English and Croatian (Table 5 and 6) range from excellent (5) to good (3) as well. None of the dialogues in this language pair

received the maximum number of three excellent (5) grades from each evaluator. In fact, the grade excellent (5) occurs considerably less frequent than for the translations in the language pair English and German. Much more often does the grade very good (4) occur. The dialogue *At the store* was graded as good (3) two times for adequacy, while the dialogues *At the hotel* and *At the bank* each were graded as very good (4) for adequacy. The average grade for translations of the dialogues *At the bank* and *At the hotel* is very good (4) for fluency and adequacy, obtaining minimum professional standard quality of the translation. The translation of the dialogue *At the store* was again graded as good (3) for fluency and as very good (4) for adequacy, denoting a minimum professional standard.

Table 5: Quality Assessment results for the language pair English and Croatian

FLUENCY	**Evaluator 1**		**Evaluator 2**		**Evaluator 3**		
	Error No.	Grade	Error No.	Grade	Error No.	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**

Table 6: Quality Assessment results for the language pair English and Croatian

ADEQUACY	**Evaluator 1**		**Evaluator 2**		**Evaluator 3**		
	Error No.	Grade	Error No.	Grade	Error No.	Grade	**Average grade**
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**

Except for the fluency and adequacy division, each of the translation errors was categorized in the following subcategories: meaning shift, grammar and syntax, lexicon, spelling and typos, style and register and coherence (Figure 1). The MT made most errors by changing the meaning of the source text or due to mistranslations. Altogether fifty-six errors occur in this subcategory in the language pair English and German and seventy errors occur in this subcategory in the language pair English and Croatian. The second highest number of errors fall under the subcategory style and register with twenty-six errors for dialogues in English and German and thirty-nine errors for dialogues in English and Croatian. A similar number of errors occur in the subcategories grammar and syntax, and coherence, followed by lexical errors. The smallest number of errors are found in the subcategory spelling and typos, with only three for dialogues in both language combinations, since only those, which affect the translation quality, were taken into account.

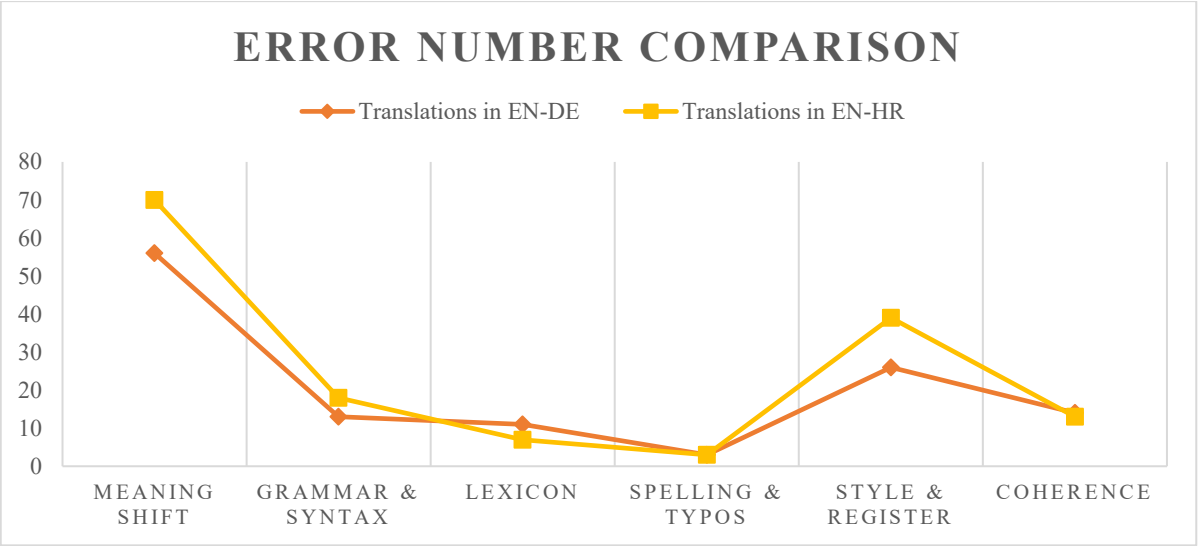

Figure 1: Comparison of error numbers in each of the categories

4.2. Post-editing

This part of the paper concentrates on post-editing as a means of evaluating the quality of the MT based on the three dialogues comprised for this research. Firstly, a short overview is given of what post-editing actually is. Secondly, the post-editing of each of the dialogues is briefly discussed.

4.2.1. What is Post-editing?

Just as translations produced by professional human translators need to be edited by a reviewer, so are machine-produced translations in need of post-editing. This is done by a linguist, ideally an expert both in the source and in the target language, who compares the translated output with the source text in order to detect any mistranslations or errors, making the translation ready for further use. How much post-editing does the machine-produced translation need depends on the quality of the MT system itself, but also on the purpose of the translation; whether the translation is meant to be published or whether it is used only for understanding a certain topic of interest makes a significant difference.

In MT post-editing, "the emphasis is on an ongoing exercise of adjusting relatively predictable difficulties, rather than on the discovery of any inadvertent lapse or error. The passages that clearly require corrections, though many of them are minor and local, are more frequent than in traditional revision" (Vasconcellos, as cited by Krings 2001:7). In other words, MT tends to make mistakes more frequently on one hand, but on the other, those mistakes can be easily predicted, especially by post-editors who are familiar with the weaknesses of MT. In addition,

Krings (2001) argues that the error differences in MT as opposed to human translation lie in frequency, repetitiveness and type. This means that human translators make mistakes less frequently than MT. Moreover, it can happen that a repeating word is once translated correctly and the second time it is mistranslated by a human translator, while MT remains consistent even in making mistranslations, producing the same mistake throughout the whole translation. When it comes to error type, translation errors in human translation can happen from a lack of concentration, fatigue or terminological misunderstanding. While resistant of these kind of weaknesses, machine translation fails in understanding context, common sense and other means of illocution.

Because of the fact that MT develops rapidly achieving better results every day, post-editors often expect the same level of quality of MT that they expect from HT. This however, is not yet possible, but many argue that machine translation will soon become so good as to replace human translators, who will therefore adjust their expertise to post-editing only.

4.2.2. Post-editing Results

In this chapter, each of the dialogues will be discussed separately. A table with the source text, the ILA-produced MT and the post-edited section will be given for each dialogue, as well as for each of the two language combinations. The purpose of the ILA is to ensure dialogue partakers to understand each other. For this purpose, only light post-editing was done, not taking into account punctuation and other style errors not influencing the understandability of the translation.

Table 7: 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 (Table 7) is the only post-edited section translated from German into English. The slight inconsistency in comparison of the MT with the source text is in number: the source text implies *Kontos*, plural and the translation says *account*, singular. The other

two examples where post-editing intervention was necessary are translations from English into German. The sentence *Sure, let me do that for you now*, caused some problems for machine translation. Firstly, since the dialogue is held at the bank, the bank official should address the customer with respect that means that *you* should be translated as *Sie*. In addition, the adjective *sure* in the translation was misplaced. 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 8: 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~~ žiro 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 language combination English and Croatian (Table 8), four sections were post-edited. Two of the four sections are translations from Croatian into English, and they are marked 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 form in singular and plural. From the context, it becomes clear that it refers to two types of accounts. This shows that the machine functions well within sentence boundary, but not beyond it. Another error in the same sentence is the indefinite article, although the definite article would be correct. The second example posed some serious problems to the machine since the translation is neither correct nor complete. *Let me do that for you* was partially translated literally and the rest of the translation was simply left out. MT also had some problems with the following example. The most serious error is the mistranslation of *a checking and a savings account*. The translation mentions checks and only a savings account, when it should say

žiro i štedni račun. The last example displays another literal translation. The more suitable translation for *susretljiv* would be *helpful*, rather than *accomodating*.

Table 9: At a 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~~ ~~du~~ Ihr bleiben

In the dialogue at a hotel, in the language combination English and German (Table 9), only two minor mistakes appear. The first is a translation from German into English, where *Bad* is translated as *bathroom* instead of *bath (tub)*. This is visible from the context, since the only room that the concierge can offer later on is one with a shower. The second error is again connected to the grammatical number. Since two people came to the hotel, the concierge should be referring to them in the plural form *Ihr*, not *du*.

Table 10: At a 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 bi ~~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

Interestingly, in this dialogue, the intervention was needed only in the segments translated from English into Croatian (Table 10). In the first example, the collocation *let me check* was not literally translated, anyways in the Croatian translation a constituent is missing. The error in the next sentence, just as in the dialogue before, lies in the grammatical number. The third example

shows MT to have problems with translating questions. *May I* was translated with *izvolite*, instead of with a polite inquiry *mogu li*. This is why the rest was translated as a declarative statement *vidim vašu osobnu iskaznicu* instead of a question. The fourth error is a mistranslation of the collocation *dial 0*, which should read *birajte 0* in Croatian.

Table 11: 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 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 11), 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 segment, it is about collocations identifying quantity: *packet of eggs* and a *bottle of milk*. The next translation, the pleasant inquiry *ich möchte* was left out, so in the post-editing *I would like to have* was added. In the third example the wrong collocation was used, namely it should read *das wäre 25 £*. Also, the word *receipt* was mistranslated with *Quittung,* which does not fit the register and should simply be *Rechnung*.

Table 12: 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 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

In the dialogue at the store led in English and Croatian (Table 12), intervention was needed in four segments. Two of the errors appear in the translation from Croatian into English and two of them in the combination vice versa. 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, missing the identifier. The translation of the second segment has a minor error, because *cube sugar* is in Croatian *šećer u kocki*. The following example needed intervention simply 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.

4.3. Automated Translation Metrics

Another tool for evaluating MT is automated translation metrics. In this section, ATM will be briefly introduced and afterwards, one of the automated metrics will be applied on the ILA-produced translations.

4.3.1. What is Automated Translation Metrics

Automated Translation Metrics refers to evaluation of machine-produced translations using automated metrics. Some of the most common are BLEU, NIST, METEOR, TER and CharacTER. Most of the automated metrics evaluate the translated content based on a similarity method. Panić (2020) explains that they compare the MT output to a human-generated reference translation on a segment level – the unit of comparison can, for example, be also a word. Besides, automated metrics use n-grams to calculate the precision scores. Automated translation metrics emerged as an answer to the question of how to conduct an objective, consistent, quick and cost-effective assessment of translations generated by MT, previously done manually by human linguists or translators. The automated metrics used for evaluating ILA-produced MT is BLEU. The BLEU (Bi-Lingual Evaluation Understudy) Metric was for the first time proposed in the 2002 paper: *BLEU: A Method for Automatic Evaluation of Machine Translation* (Papineni et al.). Since it works on a similarity-based method, BLEU does not measure the overall quality of a translation. Rather, BLEU measures adequacy of MT by comparing the word precision and fluency of MT by calculating the n-gram precision. The translation score is given on a scale 0-100, with 100 representing a 100% match of the MT translation output with the reference translation. The creators of the 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).

Image 4: BLEU comparing sentences

4.3.2. BLEU Results

In this chapter, the results of the BLEU metrics of the ILA-produced MT output will be discussed. Each of the translated dialogues underwent a light post-editing. The post-edited translation will for this purpose serve as the reference translation. Table 13 gives an overview of the BLEU results.

Table 13: Overview of BLEU metrics results

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

It should be noted that in each of the dialogue situations, the one in the language pair English and German received a higher score. The English and German dialogue *At the bank*

resulted with 93.18, while the same dialogue in English and Croatian scored 80.06/100. In the second situation, *At the hotel*, the dialogues in both language pairs gained the highest score of all the three. The dialogue in English and German scored 96.32/100 and the dialogue in English and Croatian a great 89.34/100. The third dialogue situation received the lowest score in both language combinations. The one in English and German received 80.27 points and the one in English and Croatian the lowest score of 78.82/ 100. Since none of the ILA-produced translations scored lower than 70, this shows an exceedingly successful MT output.

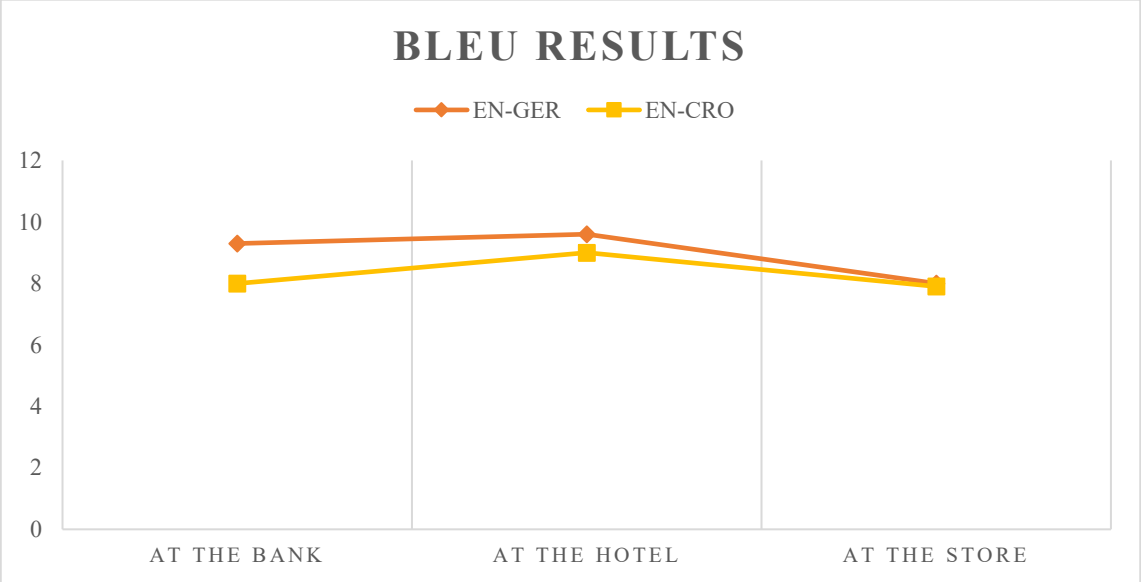

Figure 2: BLEU results for each language combination

Figure 2 above shows the BLEU results for all three dialogues in each of the language pairs. The average score of all three dialogues in the language combination English-German amounts 89.92/100. For the language combination, English-Croatian the average score amounts 82.74/100. As expected, the English-Croatian dialogues scored lower, but the difference is surprisingly small.

4.4. Comparing MT with HT

This chapter focuses on the comparison of MT with HT based on the same dialogues produced by the ILA app and by fellow students, graduates of the MA in Translation Studies in Osijek. Two students translated the source dialogues into the language pair English-German and two students translated the source dialogues into the language pair English-Croatian.

4.4.1. MT vs. HT Methodology

For the comparison of machine-produced and human-produced translations of the dialogues, a basic translation-marking grid was adjusted to meet the needs of this research. The adjusted translation marking grid contains five categories, each of which is used as one of the constituents in grading MT as well as HT. These overlapping categories are: Meaning, Grammar, Register, Clarity and Addition/Omission. The category *meaning* encompasses all distortions affecting the understandability of the text. *Grammar* includes misuses of tense or mood giving rise to interpretations other than the one intended in the source text. *Register* implies inappropriate register in the specific situation affecting the translation flow. Lack of clarity affecting the readability of the text falls under the category *clarity*. And lastly, each addition or omission altering the meaning of the text is marked in the category *addition/omission*. For each error in the category *meaning* and *addition/omission* two points are given, since those errors affect the translation quality the most. For the remaining error categories, one point per error is given. The overall translation quality is highest at zero points.

Table 14: MT vs HT translation marking grid

Abbreviation	**Type of error**	**No.**	**Points**
SENSE	Meaning (2)		
GR	Grammar (1)		
REG	Register (1)		
CL	Clarity (1)		
ADD/OMISS	Addition/Omission (2)		
	Total points		

The errors are marked in the translations using abbreviations given in Table 14 above. Each of the dialogues was reviewed separately; the results are presented in tables and compared.

4.4.2. MT vs. HT Results

As visible in the Table 15 below, the ILA-produced translation of the dialogue *At the bank (EN-DE)* in the language pair English-German received eight error points. The errors occur in each of the given category, while two of them occur in the category *Register*. The first student-produced translation received five error points with two errors in grammar, one mistranslation and one inappropriate register. The second student translation received only three

error points. With only one error in grammar and one omission, it is the most successful of the three translations.

Table 15: Error points for the dialogue *At the bank (EN-DE)*

At the bank (EN-DE)		ILA TRANSLATION		STUDENT TRANSLATION 1		STUDENT TRANSLATION 2	
Abbreviation	**Type of error**	No.	Points	No.	Points	No.	Points
SENSE	Meaning	1	2	1	2		
GR	Grammar	1	1	2	2	1	1
REG	Register	2	2	1	1		
CL	Clarity	1	1				
ADD/OMISS	Addition/Omission	1	2			1	2
	Total points		**8**		**5**		**3**

Table 16 shows that the translation of the dialogue *At the hotel (EN-DE)* produced by ILA received nine error points, student translation 1 received seven error points and student translation 2 received four error points. The MT had again most problems with register, followed by mistranslations, one omission and a minor grammatical error. Surprisingly, the first student translation had two errors for wrong register and also two mistranslations. The second student translation is the most successful one with minor errors regarding meaning, grammar and clarity.

Table 16: Error points for the dialogue *At the hotel (EN-DE)*

At the hotel (EN-DE)		ILA TRANSLATION		STUDENT TRANSLATION 1		STUDENT TRANSLATION 2	
Abbreviation	**Type of error**	No.	Points	No.	Points	No.	Points
SENSE	Meaning	2	4	2	4	1	2
GR	Grammar	1	1	1	1	1	1
REG	Register	2	2	2	2		
CL	Clarity					1	1
ADD/OMISS	Addition/Omission	1	2				
	Total points		**9**		**7**		**4**

The ILA-produced translation of the dialogue *At the store (EN-DE)*, as visible in Table 17, received again nine error points, followed by student translation 2 with four error points and student translation 1 with only two error points. This time, the most problematic categories for the MT were grammar and clarity with three errors in each category. There was also one error in

register and one mistranslation. Student translation 2 contains only one mistranslation and one omission. Student translation 1 is the most successful translation with only two grammatical errors.

Table 17: Error points for the dialogue *At the store (EN-DE)*

At the store (EN-DE)		**ILA TRANSLATION**		**STUDENT TRANSLATION 1**		**STUDENT TRANSLATION 2**	
Abbreviation	**Type of error**	No.	Points	No.	Points	No.	Points
SENSE	Meaning	1	2			1	2
GR	Grammar	3	3	2	2		
REG	Register	1	1				
CL	Clarity	3	3				
ADD/OMISS	Addition/Omission					1	2
	Total points		**9**		**2**		**4**

Error points for translations of dialogues for the language pair English-German range from a maximum of nine to a minimum of two error points. For each of the dialogues, MT produced translations with the most errors (Figure 3). The highest number of errors produced by MT was found in the dialogue *At the store*, while the HT of the same dialogue proved to be most successful with the lowest number of error points.

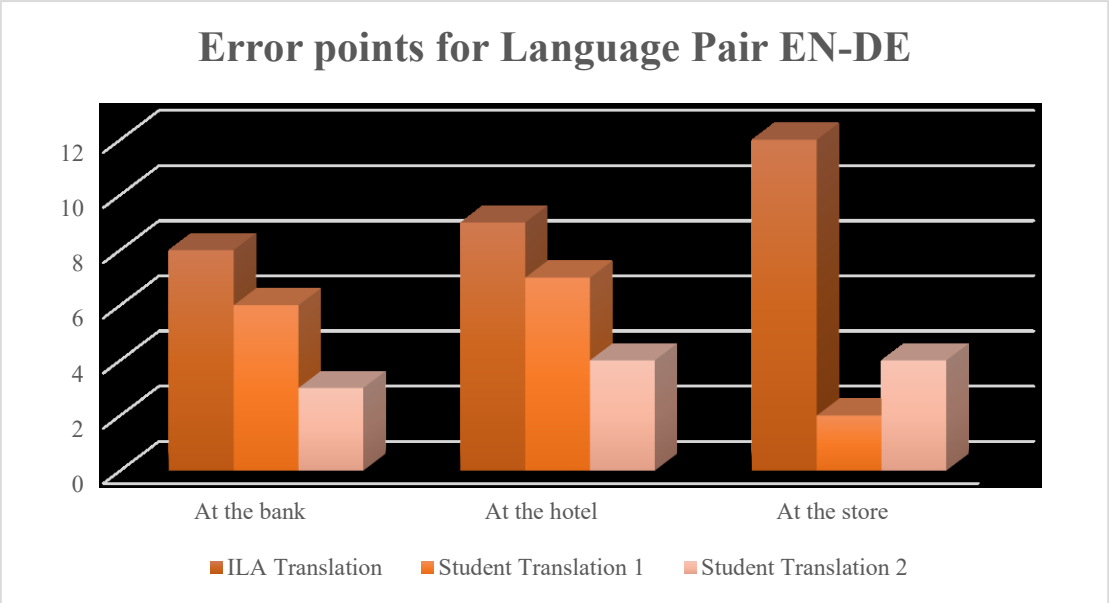

Figure 3: Review of error points

As visible in table 18, the translation of the dialogue *At the bank (EN-HR)* produced by ILA contains thirteen error points, followed by student translation 1 with seven error points and

student translation 2 with five error points. MT had most difficulties with the usage of inappropriate register and with clumsy translations. There were also two mistranslations, one grammatical error and one superfluous addition. Student translation 1 contains three grammatical errors and two mistranslations. The most successful translation, student translation 2, contains one mistranslation, one grammatical error and the register usage was inappropriate two times.

Table 18: Error points for the dialogue *At the bank (EN-HR)*

At the bank (EN-HR)		**ILA TRANSLATION**		**STUDENT TRANSLATION 1**		**STUDENT TRANSLATION 2**	
Abbreviation	**Type of error**	No.	Points	No.	Points	No.	Points
SENSE	Meaning	2	4	2	4	1	2
GR	Grammar	1	1	3	3	1	1
REG	Register	3	3			2	2
CL	Clarity	3	3				
ADD/OMISS	Addition/Omission	1	2				
	Total points		**13**		**7**		**5**

Table 19 shows that the ILA-produced translation of the dialogue *At the hotel (EN-HR)* contains ten error points with as many as four errors in grammar, two literal translations and two errors for register. The second student translation received six error points containing two errors in grammar and register, and one literal translation. The first student translation received only one error point for wrong register usage and is therefore the most successful of all the three translations.

Table 19: Error points for the dialogue *At the hotel (EN-HR)*

At the hotel (EN-HR)		**ILA TRANSLATION**		**STUDENT TRANSLATION 1**		**STUDENT TRANSLATION 2**	
Abbreviation	**Type of error**	No.	Points	No.	Points	No.	Points
SENSE	Meaning	2	4			1	2
GR	Grammar	4	4			2	2
REG	Register	2	2	1	1	2	2
CL	Clarity						
ADD/OMISS	Addition/Omission						
	Total points		**10**		**1**		**6**

As visible in Table 20, the translation of the dialogue *At the store (EN-HR)*, produced by ILA received nine error points with as much as four clumsy translations and inappropriate

collocations, three grammatical errors and one mistranslation. Student translation 2 received four error points with one mistranslation and two clumsy translations. Student translation 1 received only three error points, being the most successful translation with one mistranslation and one grammatical error.

Table 20: Error points for the dialogue *At the store (EN-HR)*

At the store (EN-HR)		ILA TRANSLATION		STUDENT TRANSLATION 1		STUDENT TRANSLATION 2	
Abbreviation	**Type of error**	No.	Points	No.	Points	No.	Points
SENSE	Meaning	1	2	1	2	1	2
GR	Grammar	3	3	1	1		
REG	Register						
CL	Clarity	4	4			2	2
ADD/OMISS	Addition/Omission						
	Total points		**9**		**3**		**4**

Error points for translations of dialogues for the language pair English-Croatian range from a maximum of thirteen to a minimum of one error point (Figure 4). Again, MT produced the most errors for each of the dialogues. Most translation errors produced by MT were found in the translation of the dialogue *At the bank.* The most successful HT proved to be the translation of the dialogue *At the hotel*, with only one error point.

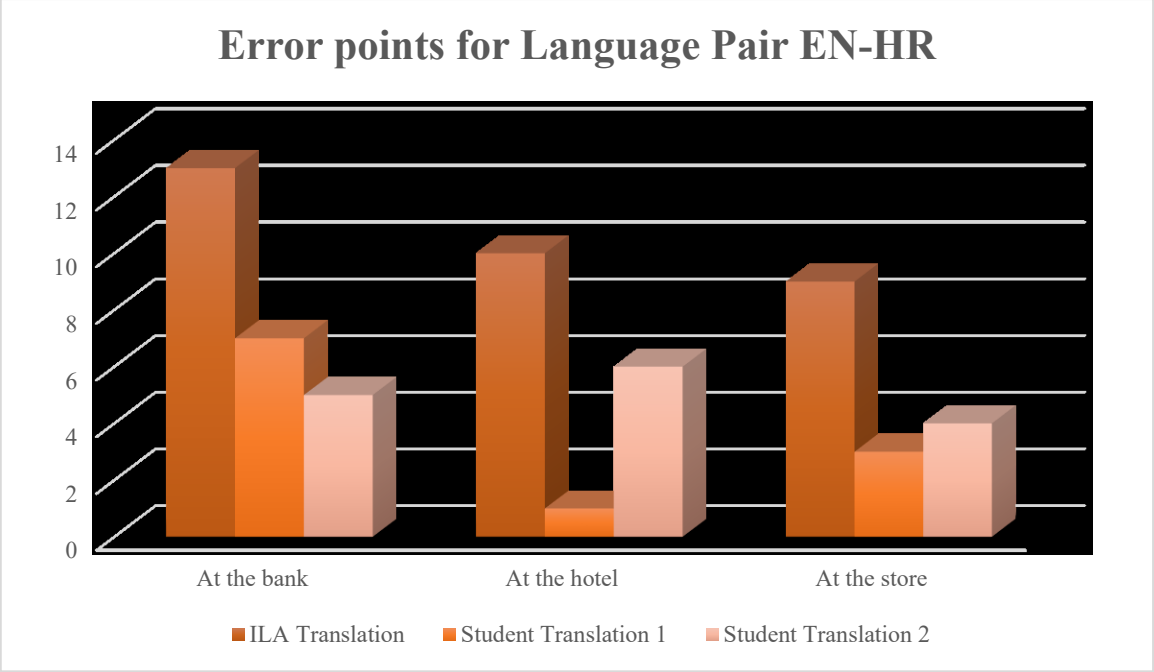

Figure 4: Review of error points

4.5. Comparing Research Results

In this part of the paper, the results of all of the previously conducted research steps will be compared based on the language pair and based on the dialogue situation. The aim of this comparison is to bring the results of the quality assessment, post-editing, the automated translation metrics and the translation marking grid together in order to see whether the language combination or the dialogue situation affect the quality of the MT. The tables below provide an overview of the research conducted on the Instant Language Assistant's output.

4.5.1 Comparing Results with Regards to Language Pair

When comparing the results (Table 21) of the quality assessment conducted by professors of translation, the average grade was calculated from the grades given for fluency and adequacy. The translation output in English and German was graded excellent (5) two times, and once very good (4). Interestingly, the translation output in English and Croatian received the grade very good (4) all three times, showing slightly lower translation quality than that in English and German. For the purpose of displaying the results of the process of post-editing the translations, the number of necessary interventions was counted, meaning that the higher the number of interventions, the lower the MT output quality. The dialogues in the English and German language pair needed an average of 4.66 interventions per dialogue. With slightly more job to do in the post-editing process, dialogues in the language pair English and Croatian demanded 8.33 interventions per dialogue. Moving on to the results of the automated translation metrics, with 100 being the highest obtainable score, the results of the BLEU metrics are given. The average BLEU score for the dialogues in English and German amounts 89.91. The dialogues in English and Croatian obtained a slightly lower average score per dialogue of 82.74. The last part of the research was the comparison of the MT with the HT in order to come up with an universal translation-marking grid for assessing the quality of the speech translation app. In addition, the higher the number of points, the lower the overall translation quality. The dialogues in English and German received an average of 8.66 points per dialogue, while the dialogues in English and German received an average of 10.66 points per dialogue.

Table 21: Comparing results with regards to language pair

Language pair		***EN-DE***	***EN-HR***
	Dialogue		
QUALITY ASSESSMENT 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
TRANSLATION MARKING GRID Number of points > MT quality	1	8	13
	2	9	10
	3	9	9

Discussing the difference in the quality of ILA-produced translations in English and German and in English and Croatian, in each of the four research categories, the dialogues translated in the language pair English and German proved to be slightly more successful.

4.5.1 Comparing Results with Regards to Dialogue Situation

The same research results are now presented having in mind the dialogue situation, that is, the topic of the translated dialogue (Table 22). The average grade of the translations of the dialogue *At the bank* is excellent (5), just as the average grade for the translations of the dialogue *At the hotel*. The dialogue at the store is graded one grade lower, namely a very good (4). When considering the post-editing of the translations, the most successful dialogue was the one *At the hotel*, which needed only 4 interventions per dialogue in each language pair, followed by the dialogue *At the bank*, which enquired an average of 6.5 interventions per dialogue. The dialogue in need of the most post-editing interventions was the dialogue at the store with 9 interventions per dialogue. Comparing the BLEU score with regards to the dialogue situation, the dialogue *At the hotel* once more proved to be the most successful one with an average BLEU score of 92.83 per dialogue in each language pair. The dialogue *At the bank* follows with an average BLEU score of 86.60. The dialogue *At the store* obtained the lowest results in this category as well with a BLEU score of 79.55. Moving on to the translation marking grid comprised for this research and in comparison to HT, the dialogue *At the bank* received an average of 10.5 points per dialogue, followed by the dialogue *At the hotel* with an average of 9.5 points. Lastly, the dialogue *At the store* received an average of 9 points per dialogue in each language pair.

Table 22: Comparing results with regards to dialogue situation

Dialogue Situation		***At the bank***	***At the hotel***	***At the store***
	LP			
QUALITY ASSESSMENT Excellent (5) to insufficient (1)	EN-DE	excellent (5)	excellent (5)	very good (4)
	EN-HR	very good (4)	very good (4)	very good (4)
POST-EDITING Number of interventions > MT quality	EN-DE	4	2	8
	EN-HR	9	6	10
AUTOMATED TRANSLATION METRICS BLEU score < 100	EN-DE	93.13	96.32	80.27
	EN-HR	80.06	89.34	78.82
TRANSLATION MARKING GRID Number of points > MT quality	EN-DE	8	9	9
	EN-HR	13	10	9

Discussing the difference in the quality of ILA-produced translations with regards to the dialogue situation, in each of the four research categories, the dialogue *At the store* proved to be the most difficult for MT, while the translation output for the dialogue *At the hotel*, proved to be the most successful one.

5. Conclusion

Professional translators, interpreters and those in need 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 quality assessment, moving on to light post-editing and automated translation metrics, and finishing off with a comparison of MT and HT produced output, this research tried to encompass all measurable components of a successful translation and thereby assess the overall quality of the S2S translations.

Before moving onto the quality assessment of the translation itself, the outcomes of the ASR and TSS technology used by the ILA app should be mentioned as well. The Automatic Speech Recognition technology gives satisfactory results. With a similar number of errors in recognising speech in all of the three used languages, ILA proves to produce good quality Speech-to-Text, regardless of the language. The users of the app have to slightly adjust to the ASR by speaking loud enough and as clear as possible, but when those prerequisites are met, ILA "understands" the spoken input very well. If the app shows some wrongly interpreted solutions, the dialogue partakers can repeat the mistaken phrase and avoid any further mistranslations and therefore, misunderstandings. Also, the Text-to-Speech Synthesis technology produces audio output of good quality. However, ILA does sound robotic, especially in Croatian.

When it comes to the overall quality assessment of the dialogue translations produced by ILA, the translations were graded based on a fluency-adequacy translation metrics. The average grades of the two levels range from excellent (5), indicating a translation of maximum/publication standard to very good (4), indicating a translation of minimum professional standard. 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 minimal BLEU score of 78.82, and a maximum of 96.32/100. Comparing the machine-produced and human-produced translations of the dialogue, the translations produced by ILA received a maximum of thirteen error points, while translations produced by students proved to be slightly more successful, obtaining a maximum of seven error points. Interestingly, human evaluation of the translation output matches BLEU and the post-editing effort to a great extent, confirming thereby the accuracy and conformity of all the three quality assessment techniques.

Overall, the research outcomes indicate high quality of translations produced by the S2ST app ILA. The main advantages are vocabulary 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, implications and others which help in the understanding and accurate translation of the conversational language. In other words, "MT systems suffer from not being able to anticipate context like human interpreters" (Müller et al. 2016: 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 research categories confirmed the same. Namely, the translation output in the language pair English-German acquired slightly better results than the translation output in English-Croatian. Still, with the average grade very good (4), the translations in English-Croatian were assessed as translations of minimum professional standard, not far behind those in English-German graded with excellent (5) or maximum/ publication standard. Moving on to the dependence of the quality of ILA-produced translations on the situation or dialogue topic, the findings of the conducted research are interesting. Again, all four research categories gave matching results. The dialogue *At the store* posed some serious difficulties to the MT. Being the most casual of the three dialogues, the participants use plain conversational language, which resulted in a MT output with the highest number of errors and clumsy translations. On the contrary, the MT output of the dialogue *At the hotel*, proved to be the most successful one. The participants of this dialogue stick to language conventions and hotel terminology, making it much easier for the machine to produce translations of high quality.

Further research on S2ST apps could be comprised of a detailed analysis of all the three layers of translation 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 defective in terms of taking into account categories like punctuation and capitalisation, which do not directly affect the quality of the speech translation output itself. Also, sometimes there is no clear distinction into which category an error should be marked. The TSS technology could be analysed based on the naturalness of the final speech production. Due to the extensiveness of the present research itself, the ASR and TSS technology could not be discussed here in further detail.

This research proved the MT to be highly productive, but there is still plenty of room for translation technologies including S2ST to improve. In order 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 spite of the breath-taking technological achievements in the translation industry until now, translation technologies are still met with scepticism. Considering the fact that machines will never be able to function completely the same as human beings, sceptics require stronger evidence that would persuade them to trust a machine. However, the fact is that machines are a part of our global society, making their way into every aspect of the human life. Having in mind the vast benefits of translation technologies in everyday lives, one must be ready to venture into the unknown, exploring the possibilities introduced to humankind by machines.

6. Bibliography

Arora, Karunesh, Sunita Arora, Mukund Kumar Roy (2013). Speech to speech translation: a communication boon. *Computer Society of India* 3: 207-2013

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

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

Jia, Ye, Ron Weiss (15th May 2019). Introducing Translatotron: an End-to-End Speech-to-Speech translation model. Available at: https://ai.googleblog.com/2019/05/introducing-translatotron-end-to-end.html (visited on 2nd June 2021).

Kohn, Marek (17th Feb 2019). Is the era of artificial speech translation upon us? Available at: https://www.theguardian.com/technology/2019/feb/17/is-the-era-of-artificial-speech-translation-upon-us (visited on 2nd June 2021).

Moskvitch, Katia (15th Feb 2017). The machines that learned to listen. Available at: https://www.bbc.com/future/article/20170214-the-machines-that-learned-to-listen (visited on 2nd June 2021).

Müller, Markus et al. (2016). Lecture translator speech translation framework for simultaneous lecture translation. *Proceedings of NAACL-HLT of the Association for Computational Linguistics.*

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

Seligman, Mark, Alex Waibel (2019). Advances in Speech-to-Speech Translation Technology. *Cambridge University Press* 12: 217-251.

Seligman, Mark, Alex Waibel, Andrew Joscelyne (2017). *TAUS Speech-to-Speech Translation Technology Report*. De Rijp, The Netherlands: TAUS BV.

Sperber, Matthias, Matthias Paulik (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.

Stein, Daniel (2013). Machine Translation: Past, present and future. Rehm, Georg, Felix Sasaki, eds. *Language technologies for a multilingual Europe.* Berlin: Language Science Press, 5-17.

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

8. Abstract

In our technologically advanced times, when machine translation becomes qualitatively more successful and quantitatively more productive, translation technologies gain more and more trust not only of the people in need of translation, but of interpreters and translators as well. Taking into account the numerous advantages of producing translations without human boundaries, MT is becoming a reliable solution to the challenges of a constantly rising demand in quick and cheap translations. This paper focuses on the assessment of Speech-to-Speech translation apps, which bring together all of the state-of-the-art translation technology, namely, Automatic Speech Recognition, Machine Translation and Text-to-Speech Synthesis. The aim of this paper is to encompass all measurable components of a successful translation and thereby asses the overall quality of translations of conversational language produced by the S2S translation app ILA. This multi-layered research consist of a quality assessment, light-post editing, an automated translation metrics and a comparison of machine and human produced translations. Moreover, the translation output is assessed with respect to the language pair and the situation in order to gain perspective on how they affect the translation output. The given results indicate a high quality of translations produced by the S2S translation app, showing the large potential of speech technologies. Additionally, this research is setting ground for further research in this field.

Key words: translation technology, speech technology, Speech-to-Speech translation apps, conversational language, ILA

9. Sažetak

U tehnološko razvijenom vremenu kada strojni prijevod postaje kvalitativno sve uspješniji, a kvantitativno produktivniji, prijevodne tehnologije zadobivaju sve više i više povjerenja ne samo od strane ljudi kojima je potreban prijevod već i od strane tumača i prevoditelja. Uzimajući u obzir brojne prednosti produkcije prijevoda bez ljudskih ograničenja, SP postaje pouzdano rješenje izazovima neprestano rastuće potražnje za brzim i jeftinim prijevodima. Ovaj se rad bavi procjenom aplikacija za prevođenje i sintezu govora, koje sjedinjuju najsuvremenije prijevodne tehnologije, naime, tehnologije za prepoznavanje govora, strojni prijevod i tehnologiju za sintezu govora. Cilj ovog rada je obuhvatiti sve mjerljive komponente uspješnoga prijevoda te tako procijeniti ukupnu kvalitetu prijevoda razgovornoga jezika nastalog u aplikaciji za prevođenje i sintezu govora ILA. Ovo višeslojno istraživanje sastoji se od procjene kvalitete, redakture teksta, automatske metrike za procjenu prijevoda, i usporedbe strojnog i ljudskog prijevoda. Nadalje, prijevodni je rezultat procijenjen u odnosu na jezičnu kombinaciju i situaciju, kako bismo dobili pregled njihova utjecaja na krajnji prijevod. Dobiveni rezultati ukazuju na visoku kvalitetu prijevoda nastalih pomoću aplikacije za prijevod i sintezu govora, pokazujući velik potencijal tehnologija za govor. Uz to, ovo istraživanje postavlja osnovu za daljnja istraživanja na ovome području.

Ključne riječi: prijevodne tehnologije, tehnologije za govor, aplikacije za prevođenje i sintezu govora, razgovorni jezik, ILA