MASTERARBEIT | MASTER'S THESIS

Titel | Title

The Role of Language Technology in Doctor-patient Communication

verfasst von | submitted by

lic. Diana-Andreea Codrean

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 070 354 331
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Translation Rumänisch Deutsch
Betreut von | Supervisor:	Univ.-Prof. Dragos Ioan Ciobanu PhD

Acknowledgments

First and foremost, I would like to thank my supervisor, Univ.-Prof. Dragos Ioan Ciobanu, who has always shown great enthusiasm and support for my topic, which gave me the confidence to successfully carry out this endeavour.

I would like to express my gratitude to my translation and interpreting teachers at our institute for their trust and guidance, which helped me grow and build the foundation of my professional journey.

A special thank you goes to my parents, who gave me the freedom to choose my own path and supported me in everything I wanted to achieve. Their faith in the choices I make has given me the strength and confidence to keep moving forward.

I would also like to thank all the participants in my study, without whose cooperation this work would not have been possible.

My heartfelt gratitude goes to my friends, whose support has meant more than I can express — whether it was proofreading, helping with experiments, or simply listening when I needed to vent. Their friendship and encouragement have made these years truly cherished in my heart and have carried me through to the completion of this work.

And finally, a very special thank you to my partner, who has been by my side every step of the way, sharing in the hard moments and the triumphs, and whose love, support, and laughter have been my anchor throughout.

Table of Contents

1. Introduction

Migration and language are two deeply connected topics. People who leave their country of origin face multiple obstacles in everyday life in a foreign place where an unknown language is spoken. Interpreters and translators, and more recently, language technology, have played a key role in facilitating the access of those people to key public services such as healthcare. After the fall of the Communist Regime in Romania in 1989, emigration restrictions were lifted, and many people migrated to countries like Germany, Hungary or Israel (OECD 2019: 26). Later, after Romania joined the EU in 2007, migration paths focused on Western European countries like Italy, Germany, Spain, the UK, Hungary, France and Austria, and the US (OECD 2019: 41). This movement kept growing over the years, and today Romania has one of the world's largest diasporas. According to Statistik Austria, at the beginning of 2024, Romanians were the sixth-largest population group living in Vienna, counting 43,173 (Turulski 2024). Consequently, a large group of Romanians resides in Austria, with different proficiency levels in German or other international languages.

For the people who do not speak the language of the host country very well, access to medical services is often delayed and cumbersome (Thonon et al. 2021: 2). Low language proficiency is associated with several medical dangers: pregnant women are less informed about beneficial treatments and are more likely to experience obstetric trauma; children are at a higher risk of severe medical issues, and oncology patients get fewer screenings (Thonon et al. 2021: 2).

Especially in healthcare, when it comes to life-or-death situations, effective communication is crucial. According to Hamai et al. (2024), communication errors are the main cause of medical incidents for those patients who do not speak the same language as their healthcare provider. This emphasises the need for an intermediary to break the language barrier. The involvement of interpreters in medical settings improves patient outcomes, encourages preventive advice from healthcare professionals and reduces the number of emergency visits (Thonon et al. 2021: 2).

When a professional interpreter is not available on site, other solutions for facilitating multilingual communication are guides and brochures, informal or ad-hoc interpreters or professional interpreters connected over the phone or via video conference, and general translation apps (Thonon et al. 2021: 2). According to Brandenberger et al. (2025), "lack of systematic screening for language proficiency of patients, low patient awareness of their own

right to request access to translation services, health care providers with limited training on transcultural competencies and time constraints" (Brandenberger et al. 2025: 2) are the main reasons for not using professional interpreting services in healthcare. In the absence of professional interpreters, the risks of misdiagnosis, limited emotional expression of the patients, additional stress for the untrained interpreters, and, in the case of relatives or children, problematic family dynamics and possible psychological trauma increase (Leanza 2007: 12). Reducing economic burdens on the healthcare system, difficulty of covering the increasingly linguistically diverse need for translations, and dissatisfaction with the accuracy of ad-hoc solutions, such as bilingual relatives or acquaintances, are motivations as to why healthcare professionals have started gravitating towards Machine Translation (MT) or even Speech-to-Speech (S2S) translation tools to assist communication (Olsavszky et al. 2025). As this phenomenon becomes more widespread and human lives are at stake, research should determine the viability of such tools as an alternative to human interpreters and to identify any possible associated risks.

1.1. Aim and Research Questions

The shift to translation technologies, driven by the "technological turn in interpreting" (Fantinuoli 2018: 3), the emergence of Large Language Models (LLMs) and the need to address growing translational demands raise questions about the implications of such tools for patient outcomes and confidentiality, and the performance difference between human and LLM-based machine interpreting. In the hope of discovering more about how these technologies can be used by the healthcare and interpreting communities, this thesis aims to assess the role of language technology in doctor-patient communication while focusing on quality and perceived understanding. Therefore, the two research questions posed in this thesis are:

> RQ1: How does S2S translation via ChatGPT compare to human interpreting when assessing the perceived understanding of healthcare providers and patients during medical consultations in German and Romanian?
>
> RQ2: What differences in quality emerge between human interpreters and S2S translation via ChatGPT when analysed through the framework proposed by Hamai et al. (2024) and MQM?

Starting from these research questions, the assumptions of this thesis are, firstly, that even though S2S technology constitutes a solid and quick alternative to professional human interpreters when it comes to exchanging information, the perceived understanding is higher for both the patients and the healthcare provider when interacting with a human interpreter. The second assumption is that participants would consciously modify their speech when interacting with ChatGPT compared to the human interpreter.

To contribute to a broader general understanding of how human and machine interpreting shape the communication in medical settings, the main goal of this thesis is to evaluate how the perception of understanding differs in human and machine interpreting and to investigate the quality of human and machine interpreting using two different quality assessment frameworks. While the primary focus lies on perceived understanding and quality, this thesis acknowledges trust and data privacy as important factors influencing the acceptance of S2S translation systems in healthcare settings and discusses them as complementary considerations.

The expected outcome is a comparative quantitative and qualitative analysis of the perception of understanding of healthcare providers and patients, as well as the quality of the two methods of interpreting. This provides useful insights for healthcare professionals, patients, translators and interpreters, for it enables them to make informed decisions as to when it is appropriate to use each interpreting method and to have a better understanding of the risks and benefits each one poses, but also for developers of S2S systems, by signalling them what the possible areas of improvement of current technologies are.

This thesis first addresses community and healthcare interpreting in Chapter 2, examining the role of the interpreter through the lens of the conduit and the triadic model. It further explores the particularities of doctor–patient communication, confidentiality as outlined in the interpreter's code of ethics, and the human interpreting effort. Chapter 3 then focuses on the perception of understanding in interpreted interactions. Chapter 4 shifts to machine translation and interpreting, detailing the beginnings of machine translation, the evolution and mechanism of S2S translation systems, interpreting tools, and relevant confidentiality policies. Chapter 5 presents the evaluation of translation and interpreting quality, followed by related work in Chapter 6, which situates the study within the existing literature.

2. Community and Healthcare Interpreting

This chapter is dedicated to one of the two main focal points of this thesis, namely, the human interpreting in community, and more specifically, in healthcare settings. First, there will be a description of the emergence of community interpreting as a profession, the role of the interpreter and how it is conceptualised in the conduit and triadic models, an overview of the relevant aspects of communication and interpreting in the healthcare context, alongside a discussion on ethical and confidentiality considerations, followed by a brief description of the human interpreting effort.

Community interpreting is one of the oldest professions, but it has not been treated as such for most of history. Compared to conference interpreting, which found its place as a profession in society as early as the 1950s, and even more so after the Nuremberg Trials, community interpreting was professionalised only towards the end of the 20th century (Pöchhacker 2022). Even though interlingual communication between individuals had been happening way before modern international conferences were being organised, Pöchhacker (2022) states that the first time professional standards were set for community interpreting was with the 1978 Court Interpreters Act in the US. Only long after that, the Directive 2010/64/EU on the Right to Interpretation and Translation in Criminal Proceedings established a comparable framework in Europe (Pöchhacker 2022).

Even with some standards set in place for courtroom interpreting, most of the practices in community interpreting were not regulated for a long time (Pöchhacker 2022). Due to the lack of professionalisation and the need to lay the foundation for it, conferences, such as the Toronto Critical Link, started in the mid-1990s (Pöchhacker 2022). Community interpreting is defined by Pöchhacker (1999) as "interpreting in institutional settings of a given society in which public service providers and individual clients do not speak the same language" (1999: 126). This broad definition encompasses a variety of fields, and healthcare interpreting, a sub-domain of community interpreting, also started to establish itself as a profession at the end of the 1990s (Pöchhacker 1999: 127).

At the beginning of the millennium, there were five main ways to ensure communication between a doctor and a patient who spoke different languages. According to Bischoff and Loutan (2004), these options were hiring a paid interpreter, or getting help from bilingual healthcare professionals, other bilingual staff, volunteers or relatives, such as friends or family. Since then, technology has offered MT and S2S translation as new tools for such situations.

Bischoff & Loutan (2004) conducted a study about interpreting in Swiss hospitals and psychiatric units, and showed that 64% of the healthcare facilities that answered the survey had used some form of interpreting, but even though a paid professional interpreter would be the gold standard, only 14% of the facilities used paid interpreters, 17% of them have access to a network of paid professional interpreters, and only 11% have a dedicated budget for interpreting services (Bischoff & Loutan 2004).

In its coalition agreement from 2021, the former German Government stated "Sprachmittlung auch mit Hilfe digitaler Anwendungen wird im Kontext notwendiger medizinischer Behandlung Bestandteil des SGB"[1] (SPD et al. 2021: 65), making the commitment that language mediation would be integrated in the Social Code Book V (SGB V), thus granting every patient who is insured by the state access to language services. The Federal Association of German Translators and Interpreters (BDÜ) responded in support of this initiative and proposed a model to ensure access and quality assurance in this process (BDÜ 2023). This shows that access to interpreting is still a relevant topic in healthcare, and measures are being implemented to facilitate healthcare access to people with limited language proficiency.

According to Pöchhacker (2000: 113), the city of Vienna recognised the importance of dealing with the language and cultural barrier between locals and non-German-speaking migrants in the 80s. In that sense, they started making efforts to facilitate their integration and use of social services by providing liaison interpreting services through an initiative of the Vienna office of the WHO Healthy Cities Network (Pöchhacker 2000).

2.1. Role of the Interpreter

In the process of professionalisation, scholars such as Roy (1992), Hatim and Mason (1997) , Wadensjö (1998), and Hale (2014) have explored the role of the interpreter and the position they have to take in a mediated conversation. Hale (2014) talks about two contradicting viewpoints. On the one hand, there is the robot-like impartial interpreter, who tries to find the equivalent of one word from one language for another word from a different language. On the other hand, we have an interpreter who can decide how to adapt the content and the delivery to suit the audience. She concludes that interpreters do not fall into one or the other category in real life, but implement a mix of both, depending on the situation (Hale 2014). By adopting the

[1] "Language mediation, including through the use of digital applications, becomes part of the German Social Code (SGB) in the context of necessary medical treatment" (transl. D.-A.C.)

first perspective, interpreters would resemble the way computers interpret, which will be discussed in Chapter 4 of this thesis. This spectrum of the impartial, invisible interpreter on one side, and the active, involved interpreter on the other, is a heavily discussed topic in interpreting studies and is represented through the conduit, respectively, the triadic model.

2.1.1.Conduit Model

Until the mid-1990s, the status quo for interpreting was the conduit model (Downie 2020). This model was first proposed by C. E. Shannon (1948) as a general communication framework. The original model describes a linear communication system made up of five parts: an information source, a transmitter, a channel as the medium to transmit the message, a receiver and a destination (Shannon 1948: 380–381). This model is a technical one, more complex, and centred around telegraphy and telephony communication. In a simplified manner applied to interpreting, it consists of three elements, as depicted in Figure 1:

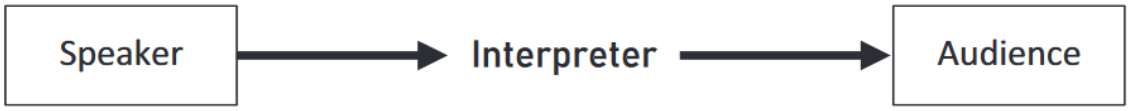


Figure 1. The conduit model adapted for interpreting (Downie 2020: 7)

As a result of interpreting scholars initially focusing on simultaneous and conference interpreting, where there usually is a speaker that addresses an audience through an interpreter, and does not necessarily get a response, the interpreter was generally considered rather a communication channel (Downie 2020). Moreso, in the early stages of interpreting studies, research was focused on the cognitive and linguistic processes, and the language transfer, rather than the social aspects of interpreting (Roy 1992).

Downie (2020) explains that by following the conduit model, interpreters become an impartial, invisible instance, with the sole function of transferring information word for word. This model nearly dehumanises the interpreter, implying that they must strive to be as invisible as possible. If this were a model that could be realistically applied in real life, it might be possible for a machine to match human interpreting performance. Even if impartiality is desired in theory, it cannot be put into practice. This model excludes considerations of interpreting ethics and disregards the role of the interpreter in decision-making.

The limitations of this model, which reduces interpreting to the mere transfer of information without context, led Seleskovitch (1968) to develop the Theory of Sense. This

theory revolves around the concept of deverbalisation and emphasises understanding and conveying meaning from the source language into the target language.

Pioneer researchers in interpreting studies like Seleskovitch (1968), Moser-Mercer (1997) and Gile (2009) focused on interpreting from a cognitive point of view and tried to create effort models to explain the process taking place inside the interpreter's mind. However, a shift emerged when scholars like Roy (1992), Wadensjö (1998), and Hatim and Mason (1997) started looking at interpreting from a social perspective. At this point, Wadensjö (1998) proposed a new model which would better suit the constellation and reality of interpreting.

2.1.2.Triadic Model

Even though settings where there is only one speaker at a time addressing an audience are usually public and common, most of the time, interpreting is needed in smaller, more private settings with at least three or more people involved, including the interpreter. Whether it is desired or not, the presence of a third person changes the dynamics of the dialogue. Even with the best intentions from all parties involved, the presence of an interpreter might unwillingly influence their behaviour, the flow of the conversation and even the outcome of the discussion (Downie 2020).

The need for interpreting usually arises as a consequence of another underlying need. First, there is a situation where two parties must communicate for a specific purpose. Then, there is the obstacle of being unable to communicate directly, which is why assistance is required (Pöchhacker 2022). Such situations need an interpreter to help the parties to achieve their communicative goal, which means the purpose of interpreting is to fulfil the purpose of the other involved parties (Downie 2020).

The fact that the interpreter is there to help someone achieve a goal leads to the assumption that they evaluate the situation and make conscious, well-thought-through choices regarding instances in which a sentence needs to be rephrased, or the listener would require additional explanation to best reach the goal of the conversation (Pöchhacker 2022). As Downie (2020: 114) states, "[g]ood interpreting is the ability to take such decisions intelligently".

Starting from a specific situation, Roy (1992) argued against the conduit model and proposed a new triadic model. In her work, she describes a situation in which multiple parties speak at the same time, and the interpreter has to actively take action to perform their duty of managing speaker turns and rendering the messages from one language into another (Roy

1992). Even though she referred to a specific case, this type of situation can occur at any time, forcing the interpreter to break their strictly neutral and passive role. This way, she states that:

> The interpreter's role is active, governed by social and linguistic knowledge of the entire communicative situation, requiring not only competence in the languages but also competence in the appropriate "ways of speaking" and in managing the intercultural event of interpreting. (Roy 1992: 21)

In the triadic model, the interpreter is seen as an active participant in the conversation. It presents a framework in which the human interpreter can actively contribute to the dialogue partners reaching their goals. However, according to Roy (1992), the responsibility of meeting the conversational goals is not entirely dependent on the interpreter, but also on the other communicating parties. If the interpreter is "as much a part of the event as the speaker and audience" (Downie 2020: 12), there needs to be a discussion on whether a machine should be part of the conversation and to what extent one can rely on it to understand an overarching goal and actively work towards it.

Horváth (2021) argues that the term interpreting must solely be associated with the human activity, and cannot be used for the S2S translation done by machines. He opts for the term automated speech translation because, like in written translation, the feedback only comes at the end, and there is no opportunity to adjust the output during the translation process. Horváth (2021) describes machine-interpreting systems as "black boxes" (Horváth 2021: 178) and compares this view of the interpreting process with the perception of interpreters' role in the conduit model, as non-participatory entities in a conversation, whose purpose is to only render the words of one language into another. For at the core of interpreting lies the pursuit of meaning and sense, Horváth (2021) clearly separates the human activity from automated speech translation.

The emergence of LLMs, however, shifts the idea of machine-interpreting systems as black boxes, because their output can be adjusted along the way through prompts, and their capacity for understanding language and reacting to feedback moves their role closer to that of interpreters in the triadic model. As Horváth (2021) states, human interpreting is embedded in the continuously evolving context of the conversation, and interpreters constantly and autonomously adapt their output, whereas machines are to this day not capable of autonomously analysing the broader context of a conversation and tailoring the result without additional intervention from the conversing parties.

Following the rationale of the interpreter contributing to the achievement of different communicative goals, Hofer-Falk (2023) presents several roles that interpreters may take in healthcare settings, depending on the expectations placed on them. Some situations require the interpreters to assume the role of co-therapists, by offering additional explanations of healthcare professionals' instructions, as well as the role of helpers, especially when it comes to sign language interpreting. They may also act as cultural mediators by providing healthcare providers with accurate descriptions of patients' states or pain levels that may be expressed differently across cultures. Interpreters may also be required to function as gatekeepers and make active decisions about how much of the original message will be conveyed into another language (Hofer-Falk 2023). Depending on their approach to cultural differences, Leanza (2007: 29) proposes an organisation of the interpreter's roles in system agents, community agents, integration agents and linguistic agents. In other words, the interpreter can either represent the dominant culture, the minority, mediate culture differences between them, or remain exclusively impartial.

Hamai et al. (2024) differentiates between positive and negative alterations done by interpreters, specifically in healthcare, and shows that when interpreters make constructive contributions to the conversation, they promote effective triadic communication, enabling all three parties – patient, physician, and interpreter – to work together towards the common goal of providing the patient with the necessary medical information. This highlights how the triadic model provides a useful framework for understanding the interpreter's contribution to collaborative sense-making in medical consultations.

2.2. Particularities of Doctor-patient Communication

The interaction between a patient and a healthcare provider is, most of the time, an expert-layperson conversation that takes place within an institutional framework, and, according to Hofer-Falk (2023), emerged as an area of study in the 70s. Usually, this communication follows a multiphase structure of mutual greeting, description of complaint by the patient, complaint investigation by the doctor, diagnosis, treatment plan and development, and closure (Hofer-Falk 2023: 38).

Hofer-Falk (2023) describes this dialogue as a sequence of questions asked by the healthcare provider and answered by the patient. The doctor is more often the one starting the conversation with a question about the patient's condition, followed by the patient's response. From there on, the doctor takes the lead in the conversation, decides on which topics they want

to focus on, and asks follow-up questions (Hofer-Falk 2023). According to Hofer-Falker (2023), this sequential change of speakers follows the process of turn-taking and creates an intrinsic coherence of the conversation. Institutional contexts, such as healthcare, are characterised by an asymmetrical distribution of interactional authority (Merlini & Favaron 2007: 107). This means that the doctor is usually the one in power to start interactions and decide when to allow the patient to speak.

The special case of interpreter-mediated medical consultations entails a different turn-taking process and shifts in the power dynamics of the interaction towards the interpreter, as the only participant who possesses knowledge of both languages. Iglesias Fernández (2010) notes, that interpreters manage turns and coordinate the conversation through verbal as well as nonverbal features. In their analysis of turn-taking patterns of this triadic constellation, Merlini & Favaron (2007: 107) deemed the following three as expected and logical in an interpreter-mediated conversation:

1. Doctor's question – interpreter's translation – patient's answer
2. Patient's answer – interpreter's translation – doctor's assessment or next question
3. Doctor's assessment – interpreter's translation – doctor's next question

In instances where these progressions are altered, takeover of interactional power by the interpreter is more clearly visible. This can happen when the interpreter asks the patient additional questions before translating their message to the healthcare provider, when they add their assessments or acknowledgements of understanding or manage overlapping speech (Merlini & Favaron 2007).

It has been known for a long time that effective and efficient communication between patients and healthcare providers improves patient outcomes (Stewart 1995). Patients who do not speak the local language receive fewer follow-up appointments, come back less often for those appointments, are more prone to not follow the doctor's instructions, make fewer comments during consultations, and the comments they do make are overlooked, thus impacting the patient outcomes negatively (Bischoff & Loutan 2004). In medical research, patient outcomes are determined by "satisfaction, compliance (adherence to treatment), knowledge, understanding, coping, quality of life/health status, recall, psychiatric morbidity (anxiety, depression), recovery" (Ong et al. 1995: 910). Iglesias Fernández (2010: 216) states that positive patient outcomes are strongly influenced by the combination of the physician's knowledge and competence and the emotional communication, i.e. doctor-patient rapport. According to Iglesias Fernández (2010), rapport is characterised by mutual understanding, a

feeling of positivity and warmth and behavioural coordination, and represents an essential component of effective healthcare, even supporting patient recovery through communication aimed to reduce anxiety.

In Norfolk et al. (2007: 690), rapport was defined by healthcare specialists as "a shared understanding or connection between practitioner and patient or client". Empathy lies at the core of rapport building, as it is built on the willingness and the ability of physicians to understand the patients' point of view, (Norfolk et al. 2007: 691). Derksen et al. (2013: 76) also highlights the importance of empathy from the perspective of patients in therapeutic care, and defines it as "the competence of a physician to understand the patient's situation, perspective, and feelings; to communicate that understanding and check its accuracy; and to act on that understanding in a helpful therapeutic way." (Derksen et al. 2013: 76)

Derksen et al. (2013: 76) define empathy on an effective, cognitive and behavioural level. The effective level refers to the attitudes and moral standards that the healthcare professionals have formed throughout their life. On a cognitive level, empathy is characterised by an "empathic skill, a communication skill, and the skill to build up a relationship with a patient based on mutual trust" (Derksen et al. 2013: 76). Through the skills of empathy and building a relationship whit the patients, practitioners aim to gain the patients long-term trust and create a favourable environment for them to share information. The communication skill aids the understanding and reflection process of the practitioner. The behavioural component includes the cognitive verbal and nonverbal skills of the physician and the affective recognition of the patient's emotional state and the consequent reaction of understanding towards them (Derksen et al. 2013: 77).

According to Derksen et al. (2013: 78), effective display of empathy in doctor-patient communication has led to "improvement of patient satisfaction and adherence, decrease of anxiety and distress, better diagnostic and clinical outcomes, and more patient enablement". This is also supported by Allison & Hardin (2020: 495), who claim that when patients feel heard, understood and cared for, they communicate personal or sensitive experiences, which can be beneficial for the medical outcomes.

Feeling understood is thus an important component in patient outcomes and is at risk when the doctor and the patient face a communicative barrier like language. Some studies also correlate the lack of interpreting services with the low utilisation of preventive medical services and examinations, and lower overall satisfaction (Bischoff & Loutan 2004). Bot (2005: 90) emphasises the importance of empathy from both the healthcare provider and the interpreter in therapeutic communication, and how pure neutrality is not the desired attitude in this context.

Moreso, Bot (2005) suggests that the interpreter plays a key role in establishing an atmosphere of trust through verbal and nonverbal communication, alongside the therapist. This expectation from the interpreters to actively contribute to building trust, establishing rapport and displaying empathy deviates from the norm of the interpreter's neutrality. Thus, Iglesias Fernández (2010: 222) found interpreting rapport as a challenge which has not been sufficiently addressed in interpreter's codes of ethics.

2.3. Confidentiality in the Interpreter's Code of Ethics

While ensuring mutual understanding is critical in medical consultations, the absence of qualified interpreters sometimes leads to untrained individuals, often patients' relatives or even children, being asked to step into this role. Such situations heighten the risk of miscommunication and breaches of confidentiality, highlighting the necessity of professional ethical standards.

While confidentiality between doctors and patients is already safeguarded by established policies and professional guidelines in healthcare, the specific context of interpreter-mediated consultations remains less clearly related. Thus, in the process of professionalization, the International Association of Conference Interpreters (AIIC) has created the first code of ethics and professional standards for interpreters (Pöchhacker 2022: 29). In healthcare interpreting, some of the first standards were set in place by the International Medical Interpreters Association (IMIA) (IMIA 2006) and the National Council on Interpreting in Health Care (NCIHC) (NCIHC 2004) and in Austria, the main institutions that guard the ethical practice of interpreters are UNIVERSITAS, with their own code of ethics, "Berufs- und Ehrenordnung" (UNIVERSITAS Austria 2017) and the Österreichischer Verband der allgemein beeideten und gerichtlich zertifizierten Dolmetscher:innen (ÖVGD) (ÖVGD 2025), which states on their website, that each member has to comply with their code of ethics and professional conduct (ÖVGD 2025), however, the code is not freely accessible.

The first three mentioned institutions consistently identify confidentiality as a fundamental ethical obligation for translators and interpreters across general and medical contexts, members being strictly bound to secrecy regarding information obtained during their work. The UNIVERSITAS code explicitly mentions that the duty of confidentiality remains in effect even after a contract has ended.

The NCIHC and IMIA codes explicitly regulate the healthcare field, offering an even clearer guide for ethical practice, the former being considerably more detailed. Interpreters are

advised to clearly communicate the scope of their role to patients, emphasising that all statements made in the presence of the healthcare provider will be interpreted. Consequently, patients are cautioned against sharing information they do not intend the provider to access (NCIHC 2004).

The NCIHC code also adds an even higher level of specificity and describes certain situations that allow for justified deviations from strict confidentiality requirements. When the well-being or dignity of the patient may be adversely affected, disclosure may be justified or even required. Furthermore, some states have laws requiring the disclosure of information regarding child or elder abuse, as well as threats of self-harm or harm to others, and in the event of legal or internal disputes, the principle of confidentiality still applies to the information provided to an arbitration court or internal board.

In addition to general confidentiality principles, the NCIHC highlights specific situations that may generate conflicts, such as patients being moved from one medical facility to another, and states that information should be kept confidential within the treatment team and can be shared with a new team only after the patient expresses written permission.

Confidentiality in healthcare interpreting involves complex ethical and practical decisions, such as determining which information should be disclosed and to whom. This proves that interpreters cannot be merely passive message transmitters but are active participants in the medical interaction. Through their personal and professional judgment, they can directly manage the conversation, safeguard patient well-being, and support high-quality care.

2.4. Human Interpreting Effort

Regardless if simultaneous or consecutive interpreting, the interpreter performs countless simultaneous operations during their activity. They need to take everything happening around them into account, from the speaker's nonverbal language to their surroundings and combine this with language skills and background knowledge to understand the source text to the best of their ability, to then produce an incredibly well-coordinated and monitored output that is also curated for the specific target audience (Downie 2020).

According to Gile (2009), the act of interpreting requires a certain amount of mental energy, consuming almost all of the cognitive resources, and when the performance requires more, the performance decreases. There are automatic operations that interpreters perform instinctively, which require minimal resources, and non-automatic operations, for which they

have to use considerably more resources to perform (Gile 2009). As they do not possess unlimited cognitive and processing ability, interpreters must manage their resources well during sessions and work towards automating more of the non-automatic operations in their interpreting skill training to maintain and increase quality.

To conceptualise the cognitive effort of interpreters in various contexts, Gile (2009) developed effort models based on the three main efforts of listening and analysis, memory and production. As this thesis focuses on doctor-patient dialogues, this corresponds to consecutive interpreting, characterised by two phases: comprehension and speech production (Gile 2009: 175).

The first phase is illustrated by Gile (2009: 175) as a sum of listening and analysis of the source language speech, short-term memory operations executed in the time between hearing the source speech and writing down the notes or storing the information in the long-term memory, note-taking, and coordination. Phase two consists of remembering, as the process of accessing the long-term memory, note-reading, production and coordination. Each of the phases encompasses all three types of efforts, with different levels of difficulty. In order for consecutive interpreting to be successful, Gile (2009: 176) presents a set of conditions that must be met at all times. The processing capacity required for listening, note-taking, memory, and coordination should not exceed the available processing capacity for each of the single elements, and the sum of their total required capacity should be lower than or equal to the total available processing capacity. Should each one of those conditions not be met, then there is a great possibility of saturation and error.

While Gile's Effort Model illustrates the substantial cognitive demands on interpreters in general consecutive interpreting, the community setting and medical context pose additional challenges that further strain mental resources. These hurdles are compounded by complex, specialised terminology, the need to manage turn-taking in triadic interactions, and the responsibility to uphold ethical standards while making real-time decisions. Moreover, the different types of cognitive efforts in human consecutive interpreting bear a resemblance to the cascading processing steps of S2S translation, which will be detailed in Chapter 4.

3. Perception of Understanding

This chapter examines how understanding is achieved and displayed in communication between interlocutors, drawing insights from discourse analysis and interpreting studies. Furthermore, it emphasises the importance of turn management in promoting understanding.

Translators, interpreters, and now translation technology serve the purpose of facilitating communication between individuals who do not speak the same language. In order for their efforts to be successful, a level of mutual understanding has to be reached. Wadensjö (1998: 24) states that establishing meaning in an interpreter mediated conversation is a common effort of all the participants of the social interaction, including the interpreter. Even though the understanding of an individual regarding a certain topic can be measured to a certain degree, individuals in a conversation cannot constantly and directly evaluate each other's understanding, so they rely on a perception of understanding, dictated by behaviours and social cues.

In cognitive science and discourse analysis, Clark & Schaefer (1989) analysed how reaching understanding in a conversation is a collaborative effort. In their discourse contribution model, when people engage in a conversation with each other, they "*contribute* to the discourse" (Clark & Schaefer 1989: 259). Being a participant in a discourse does not just mean taking a turn to speak in a conversation. To contribute to a discourse, a complex set of coordinated activities is required from each party to ensure mutual attentiveness and understanding (Clark & Schaefer 1989).

After analysing models of discourse from different research fields, Clark & Schaefer (1989) present "*Common ground*" (Clark & Schaefer 1989: 260) and "*Accumulation*" (Clark & Schaefer 1989: 260) as two of the assumptions underlying every perspective on discourse. Each person enters a conversation with a set of knowledge and beliefs, but also with certain presuppositions. Those presuppositions can be common between the participants, but they may also vary (Clark & Schaefer 1989: 260–261). The information that each participant possesses at the beginning of a conversation, as well as the presuppositions each person thinks they share with the other one in a discourse, is the "*Common ground*" (Clark & Schaefer 1989: 260). This common ground can, however, change and be redefined during the discourse. If the presuppositions one party had at the beginning are proven wrong or do not match the ones of the interlocutor, the common ground is updated with the new information. This is what Clark & Schaefer (1989) describe as "*Accumulation*" (Clark & Schaefer 1989: 260). One way of

reshaping and adding to the common ground is through assertions (Clark & Schaefer 1989: 261). If there is a preexisting presupposition – for example, if two people have a conversation and both assume about the other one that they are healthy and feeling good – that is their common ground. Suddenly, one person says that they got a migraine, and so the presupposition is not true anymore, and the common ground is updated with the additional knowledge (Stalnaker 1978: 323).

In their model of discourse contribution, Clark & Schaefer (1989) propose the notion of "*grounding*" (Clark & Schaefer 1989: 263), which is the criterion that needs to be met, in order for common ground to accumulate. Grounding is the common effort of the participants of a conversation to establish that they have understood each other to an acceptable degree for achieving their purpose (Clark & Schaefer 1989: 263).

In order to bring a contribution to the discourse and meet the grounding criterion, the people taking part in the discourse have to reach a point of mutual belief, a feeling that they have understood the others and have been understood themselves (Clark & Schaefer 1989: 264–265). If this cannot be done in one turn, it may require more turns and repairs to get to a mutual understanding. Schegloff et al. (1977) explain that "An 'organization of repair' operates in conversation, addressed to recurrent problems in speaking, hearing, and understanding" (Schegloff et al. 1977: 361). For Clark & Schaefer (1989) however, repairs are not sufficient to reach grounding and have to be complemented by positive evidence of understanding.

The recursive process of contributing to the discourse conists of a presentation and an acceptance phase defined by Clark & Schaefer (1989) as follows:

> Presentation Phase: A presents utterance u for B to consider. He does so on the assumption that, if B gives evidence e or stronger, he can believe that B understands what A means by u.
> Acceptance Phase: B accepts utterance u by giving evidence e' that he believes he understands what A means by u. He does so on the assumption that, once A registers evidence e', he will also believe that B understands. (Clark & Schaefer 1989: 265)

In the presentation phase, A must make sure to communicate as clearly as possible and make self-repairs if errors occur in his or her utterance, in order to facilitate understanding for B. This leads to the acceptance phase in which B displays positive or negative evidence of understanding. According to Clark & Schaefer (1989), there are five types of positive evidence that can be displayed by B:

1. Continued attention. B shows he is continuing to attend and therefore remains satisfied with A's presentation.
2. Initiation of the relevant next contribution. B starts in on the next contribution that would be relevant at a level as high as the current one.

3. Acknowledgement. B nods or says “uh huh,” “yeah,” or the like.
4. Demonstration. B demonstrates all or part of what he has understood A to mean.
5. Display. B displays verbatim all or part of A’s presentation. (Clark & Schaefer 1989: 267)

Clark & Schaefer (1989) present these types of evidence from weakest to strongest. When B shows A through silence and eye contact, that they are following what A is saying and do not interrupt the speaker, they display continued attention. From there on, the evidence becomes easier to investigate, starting with the initiation of the next relevant contribution to the conversation, meaning that B is accepting A’s presentation by continuing the conversation at the same informational level. B can also show understanding by acknowledging it verbally or nonverbally, through a nod, for example. Lastly, B can demonstrate to A what they gathered from the presentation by phrasing it in their own words or displaying it by repeating A’s exact words.

Since this process is described as recursive, B’s acceptance of A’s presentation becomes a presentation for A that they, in turn, can either accept or not. To avoid an endless loop, Clark & Schaefer (1989) introduced the “strength of evidence principle” (Clark & Schaefer 1989: 268). According to this,the evidence A must provide after B signals their acceptance of A’s initial presentation will be weaker than the evidence B showed in the first round. The cycle is eventually concluded with the weakest form of evidence.

In interpreting studies, Wadensjö (1998) presents two models of talk, which outline the way people comprehend language. When speaking about understanding the meaning of language in use, Wadensjö (1998) differentiates between “talk-as-text” (Wadensjö 1998: 21) and “talk-as-activity” (Wadensjö 1998: 21). In the first instance, the meaning of an utterance is given by the grammatical and syntactical framework of the language used and relies solely on the text. Morphemes which carry meaning are parts of words, and the words are part of syntactical structures, which build sentences that make up a text. The second type of sense-making derives meaning from several sources simultaneously. A text is a product of understanding the current situation, the verbal and non-verbal cues of the interlocutors and the grammatical and syntactical framework of the language (Wadensjö 1998).

Analysing language use through the “talk-as-text” (Wadensjö 1998: 21) perspective seems to be an unfitting way for an interpreter-mediated situation, for it does not include contextual information, which is vital for a correct interpretation. An interpreter takes into consideration “social and cultural conventions tied to language use in interaction generally, and language use specifically within one type of activity or the other” (Wadensjö 1998: 24) when

deciding how to translate the information. Until recently, machines were limited to the dictionary meaning of words, but according to Vaswani et al. (2017), through the development of attention mechanisms, machines became context-aware as well.

In a dialogue, meaning is also conveyed through conversational mechanisms that exceed the words in a language and are culturally dependent. In a mediated conversation, the interpreter is the only person aware of those aspects from both sides and uses that knowledge to steer the flow of the conversation, potentially influencing its outcome (Roy 1992).

Turn-taking is one of the conversational mechanisms that can be driven by the interpreter. Being the intermediary who knows both languages and cultures of the speakers, the interpreter exchanges communication turns with each speaker, usually consecutively, and plays a key role in the management of those turns (Roy 1992).

A distinct occurrence in interpreted conversations is overlapping talk. Because the most common interpreting mode in dialogue settings is consecutive, the second participant does not immediately receive the message of the first speaker, as they must wait for the interpretation. During the interpretation, however, the first speaker might further add to the conversation, thus overlapping with the interpreter. Simultaneous speech can also occur between the two interlocutors, who do not speak the same language. In both cases, the interpreter proactively steers the conversation, making use of different verbal or non-verbal communication mechanisms to manage the turns (Roy 1992). In a conversation in which the interlocutors share no common language and are also not familiar with each other's cultural cues, the effective turn management of the interpreter, who is aware of those differences, might contribute to the overall understanding between the parties.

4. Machine Translation and Interpreting

This chapter will follow the development of the second central aspect of the present thesis, MT and S2S translation. It begins by outlining the early development of machine translation, moving on to the evolution of the mechanisms of S2S translation systems, and then the interpreting tools currently on the market. Finally, this chapter addresses issues related to confidentiality policies and ethical considerations associated with the use of translation tools.

4.1. Beginnings of Machine Translation

Even though the idea of a device that can break language barriers had been around for many years, the first feasible concepts appeared in 1933, when two independent patents for similar translation machines were issued in France to Georges Artsrouni and in Russia to Petr Trojanskij (Sin-wai 2023: 23). The two devices were designed to be used in the translation process as mechanical dictionaries (Sin-wai 2023).

After the end of World War II, the new electronic digital computer allowed for further development of the translation machines. In 1946-47, the British crystallographer Andrew Booth and Warren Weaver, the vice president of the Rockefeller Foundation, initiated the first ideas of a machine translation system (Sin-wai 2023). In the next two years, Booth and Richard H. Richens used their expertise and made individual efforts to face the challenges of morphological analysis and ambiguity, in the hopes of creating a mechanical dictionary (Hutchins 2023).

In 1954, Léon Dostert and Paul Garvin from Georgetown University worked together with IBM and made the first demonstration of an MT system that translated from Russian into English, using only a limited vocabulary and a small number of grammar rules (Sin-wai 2023). This sparked the interest of the science community and increased funding, which led to 5 other countries (Russia, the United Kingdom, Japan, China and Czechoslovakia) developing their own MT systems (Sin-wai 2023). The Georgetown project was, however, cancelled in 1963, and one year later, the US government brought together seven experts to research MT and thus established the Automatic Language Processing Advisory Committee (ALPAC) (Sin-wai 2023).

ALPAC issued a report in 1966, which further discouraged the funding of research in MT, because the development of a well-performing MT was highly unlikely in the near future, and at that time, and not cost-efficient (ALPAC 1966). Although there was limited progress in

the years to come, Peter Toma, who previously worked for Georgetown University, managed to create a MT system called SYSTRAN, which is still used to this day (Sin-wai 2023).

The most notable events from the end of the 70s and during the 80s were the rise of the MT systems TAUM-METEO and EUROTRA, the former developed in Canada for translating weather forecasts, and the latter used by the European Economic Community (Sin-wai 2023).

Afterwards, driven also by the evolution of Artificial Intelligence (AI), several other systems like TRANSLATOR, ATLAS/I and ATLAS/II, ATHENE, METAL, TranStar, BehaviorTran, SUSY, BULTRA, and others, emerged for different purposes, supporting different language combinations (Sin-wai 2023).

Up until the end of the 80s, MT systems relied on "morphological, syntactic, semantic, and contextual knowledge about both the source and the target languages, respectively, and the connections between them" (Xiaojing & Shiwen 2023: 193) in order to translate. This model became known as Rule-Based Machine Translation (RBMT). The system was reliable, but since language is always evolving, it is difficult to keep it up to date, and so it soon became overshadowed by the new corpus-based Statistical Machine Translation (SMT), which uses machine learning and parallel texts to determine the most statistically probable translation for elements, that is, words, phrases or syntactic constructions, of the texts (Xiaojing & Shiwen 2023). Although the idea of a statistical model emerged in 1949 with Warren Weaver's Memorandum regarding the possible future of MT, it only became a reality in 1991 (Liu Yang & Min 2023). For the next few years, the best performance was achieved by combining RBMT and SMT, until the 2010s, when the most recent system, Neural Machine Translation (NMT), became the state of the art. The integration of deep learning and neural networks in MT opened the door to new performance standards in the world of translation (Xueting & Chengze 2023).

Even if the technological developments in MT started as early as 1933, most of the people outside of the research community knew very little about them (Jin Yang & Lange 2003). The rise in popularity of machine translation started in 1997 with Bablefish, a website that offered free online translation services, from AltaVista and Systran Software Inc. (Jin Yang & Lange 2003). The revolutionary technology first offered rapid translation into ten European languages and gained traction, reaching 1.3 million translations per day in 2000 (Jin Yang & Lange 2003). This marked the beginning of people resorting to computers for their translation needs when it came to written texts.

After a paper on the merging of neural networks and machine translation was published in 2013, Google started investing in this idea and launched an online neural machine translation system (GNMT), which supported 8 language pairs in 2016 (Turovsky 2016) Soon after, the

same company proposed a method of training multilingual NMT models to translate between various language pairs within a single model enabling zero-shot translation (Johnson et al. 2017: 2–3). This way, a model can translate between language pairs it has never seen together before (Johnson et al. 2017). Such a model could, for example, translate between Romanian and Hungarian if it is familiar with Romanian-Italian and Italian-Hungarian. This development meant that models could become scalable quickly.

Because NMT models are trained on bilingual corpora, they learn patterns from those pairs of translated sentences, and the more training data they have, the better they get at producing more accurate and fluent translations. The source language text is first processed as a whole into contextual vectors by an encoder and then translated into the target language, one token at a time, by a decoder (Zeng & Liang 2024: 1). Vaswani et al. (2017) presented the next development – the transition to transformer-based architectures, which do not function sequentially, but rather analyse all the tokens at once, making simultaneous connections through a mechanism called attention. In a simplified manner, in a transformer-based translation system, every element of the input assigns an attention score to itself and to the others and, through a parallel multi-layered process, determines the important elements from the input (Vaswani et al. 2017).

However, according to Zeng & Liang (2024) NMT models face challenges with regard to the immense costs of creating large bilingual corpora for training and the task load of fine-tuning the model for specific domains or purposes. To improve the quality of the translated output, the models needed to gain a broader understanding of how language works and flexibility in its use. With this goal and alongside the foundation of the transformer-based architectures, LLMs came to be the most recent advancements in translation technology. Zeng & Liang (2024) define an LLM as "a pre-trained model that uses an unsupervised machine learning method to train a deep neural network, usually with hundreds of billions of parameters, on large unannotated data" (Zeng & Liang 2024: 1). These models are not only capable of translating oral or written input into a large number of languages, but also perform other natural language processing tasks such as searching the internet, answering questions, making conversation, generating text role-play and many more (Zeng & Liang 2024: 1).

4.2. Evolution and Mechanism of Speech-to-Speech Translation Systems

S2S translation, also called Machine Interpreting (MI) or Automatic Speech Translation (AST), enables two people speaking different languages to have an oral conversation, without minding the language barrier (Tan 2023: 726). A classical S2S translation system consisted of three main components (as depicted in Figure 1): the Automatic Speech Recognition (ASR), the MT and Text-To-Speech (TTS) synthesis for spoken output (Tan 2023).

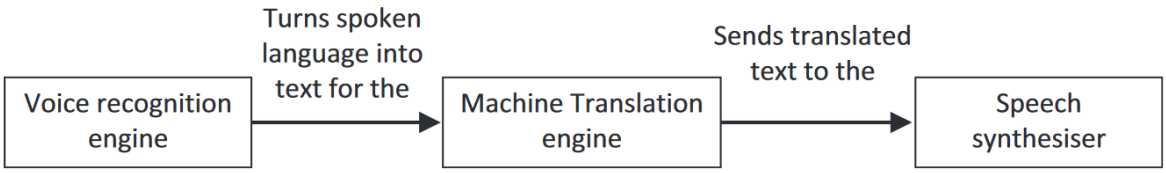


Figure 3. Model of Machine Interpreting (Downie 2020: 37)

The history of S2S translation started more than 30 years ago, but it is still shorter than that of other translation technologies (Sin-wai 2023: 36). Even though, up until recently, the general public could not benefit from technology to translate spoken language, efforts have been made since the 1970s to create such tools (Kitano 1994). The most notable research projects, however, started in Japan at the Advanced Telecommunication Research Institute International (ATR), the University of Karlsruhe in Germany and the Carnegie Mellon University (CMU) and IBM Research in the US at the end of the 1980s – beginning of the 1990s (Tan 2023).

The first system, which opened the door for the possibility of speech-to-speech translation in 1989, was ΦDMDIALOG (Kitano 1994: 1). The system was able to translate between English and Japanese, using the Advanced Telecommunication Research Interpreting Telephony Research Laboratories domain (Kitano 1994: 1). The next milestone was JANUS, a multi-parser system from the CMU, which worked from English into Japanese and German (Waibel et al. 1991).

Achieving usable spoken language translation is an immense task. S2S translation entails the following:

> The task of speech-to-speech translation ultimately requires recognition and understanding of speaker-independent, large vocabulary, continuous speech in the context of mixed-initiative dialogues. It also needs to accurately translate and produce appropriately articulated audio output in real-time. (Kitano 1994: 1)

According to Kitano (1994), there were two types of challenges that S2S translation faced in 1994, namely speech recognition and translation. Speech is rarely prepared beforehand or read

out loud; it is not always coherent and fluent, it naturally contains interruptions, mispronunciations or emphasis on specific words or syllables, and it can be disrupted by the noise of the environment or countless other external factors (Ji et al. 2023). In other words, to reach a high performance at delivering a spoken translation of a text, S2S translation systems have to first reach a very high performance at recognising natural human speech. Kitano mentioned "vocabulary size, speaker-independency, continuous speech and task characteristics" (Kitano 1994: 4) as the obstacles of speech recognition at the time.

Speaker-independency is the most difficult one, as no two people speak the same way. In the early 2000s, ASR was mainly still speaker-dependent, and systems were trained with multiple domain-specific recordings of the same person, accurately transcribed (Ji et al. 2023). These recordings were divided into segments. This way, the system could predict the next most likely continuation for a text, and even reach an accuracy of above 90% (Ji et al. 2023). As a result of the processing power of computers gradually increasing from that point on, and the recording databases becoming larger, speaker-independency on a large scale was reached soon after (Ji et al. 2023).

Between 1992 and 2000, the German Federal Ministry for Education and Research funded one of the earliest examples of a speaker-independent S2S translation system called Verbmobil (Tan 2023; Wahlster 2000). Available on mobile devices and focusing on three business domains, it could recognise speech, translate and audio-produce the output, supporting German, English, and Japanese (Wahlster 2000).

In the 2000s, the US military Defense Advanced Research Projects Agency (DARPA) worked together with various international research laboratories and contributed to the development of S2S translation through three major projects (GALE, TRANSTAC, and BOLT) (Tan 2023). These were meant to be used for US military operations in Arabic-speaking countries (Tan 2023).

S2S translation tools up to the end of the 2010s were still using the cascade model of ASR-MT-TTS, but, at the end of the 2010s, neural automatic speech recognition became the state of the art. As NMT described in Chapter 4.1, neural speech recognition also learns patterns within the input content and generates the most likely output based on that. Because of their reliance on sequences of patterns, neural architectures designed for sequences, such as recurrent and convolutional architectures, were used at first (Ji et al. 2023).

The next technological advancement was to skip the intermediate step of MT and translate the input speech directly into the output speech. Jia et al. (2019a) stated that end-to-end models, which did not rely on breaking the interpreting process into a cascade of steps,

performed faster and produced acceptable outputs. Soon after, in 2019, Google announced their new system, Translatotron which is able to do a direct speech-to-speech translation with a sequence-to-sequence model, skipping the transcription and written translation part altogether and improving latency compared to the cascade models (Jia et al. 2019a).

This works as follows and as depicted in Figure 4: Translatotron recognises the source speech, creates spectrograms of the sound and then translates that into spectrograms of the target language content. Additional features like the vocoder and the speaker encoder make the produced audio content sound more natural and even copy the voice of the speaker, to sound as if they were speaking in a foreign language (Jia et al. 2019b: 1124).

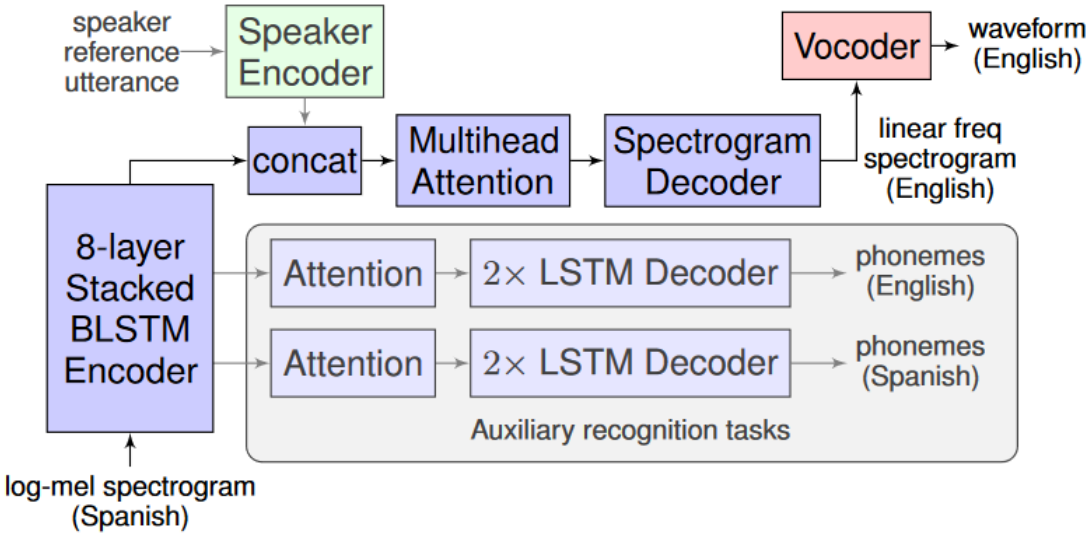


Figure 4. Architecture of Translatotron (Jia et al. 2019: 1124)

Even though this was a remarkable breakthrough, the article concluded with the fact that end-to-end models could not yet reach the accuracy and performance of cascade models and had to be further improved to replace the technology at that time (Jia et al. 2019a).

Soon after, in 2022, came Translatotron 2, which "consists of a speech encoder, a linguistic decoder, an acoustic synthesiser, and a single attention module that connects them together" (Jia et al. 2022: 10120). The improvements compared to the previous model are in terms of robustness, prosody and voice preservation, as well as privacy and security. Translatotron 2 uses an integrated speech encoder that prevents the misuse of the speaker's voice or voice cloning (Jia et al. 2022). Despite its impressive performance and improvements, there are still some downsides to this model, such as the massive amounts of training data and the computational power it requires. Therefore, cascade models still represent the most effective approach. However, as each step relies on the previous, there are several possible breaking points in the process.

4.3. Interpreting Tools on the Market

Based on the decoder of the new transformer architecture of artificial intelligence discussed in Chapter 4.2, OpenAI (Open Artificial Intelligence) launched in 2022 its chatbot ChatGPT (Chat Generative Pre-Trained Transformer) (OpenAI 2022). Before becoming a global phenomenon, GPT-1 started in 2018, by initially being pre-trained on unlabelled data, and only afterwards being fine-tuned with labelled data (Hua et al. 2024). Unlike the natural language models at the time, which were trained using labelled data, GPT learned to generalise from the unlabelled data. For the next generations, GPT-2 and GPT-3, the datasets were progressively scaled up from corpora like Books1 & Books2, WebText2, CommonCrawl, Wikipedia, etc. and in 2020, GPT-3 was able to perform most natural language processing tasks (Hua et al. 2024). This was then improved by an even more fine-tuned model called InstructGPT, using Reinforcement Learning from Human Feedback, which then led to the emergence of the chatbot ChatGPT (OpenAI 2022).

Since then, ChatGPT has been constantly improved and is able to hold conversations, summarise content, tell stories, engage in role-play, translate and much more. It understands computer languages and even not widely spoken languages and has knowledge in almost any field (Hua et al. 2024). As of April 2026, the newest version available is GPT-5.4 Thinking (Open AI 2026).

Following the technological milestone of NMT in the 2010s, countless companies have launched automatic translation tools and apps, each of them designed for different purposes and excelling in different domains. In terms of classical NMT translation systems, the current most well-known tools are Google Translate (Google 2025b) and DeepL (DeepL SE 2025). Although some studies show that DeepL has better quality when translating written texts in certain language pairs than Google Translate (Deng & Fan 2025; Sabrina et al. 2025); the former does not offer a direct interpretation function, but can only take voice input from the source language, translate it into text, and then read it out loud via speech synthesis in the target language. To carry on a conversation, the source language must then be changed at each speaker's turn, making it a more cumbersome process to have a multilingual dialogue.

As for the newly popularised LLMs that are capable of translating between multiple language pairs, two of the leading ones are Gemini (Google 2025a), formerly known as Google Bard AI, and ChatGPT (OpenAI 2025a), both offering consecutive interpreting functions in several language pairs. As translation based on LLMs is a relatively recent innovation, there is

still an ongoing debate on the performance of such systems against NMT tools like Google Translate and DeepL.

Putera & Sujana (2024) compared Google Translate with Bard when it comes to translating abstracts from scientific journals and discovered that Google Translate had a slightly greater accuracy than Bard. Hidayati & Nihayah (2024) analysed Google Translate, GPT-4 and Bard in terms of student preferences, especially those who do not study English, regarding the translation of abstracts from international journals, and the translations made with ChatGPT and Bard were perceived as having greater accuracy, sensitivity to context, and more natural and coherent translations. Li (2024) assessed the quality of ChatGPT-3.5 and DeepL, amongst other tools, in translating from English to Chinese and vice versa, and for this language combination, the LLM showed no significant refinement, and all of them performed better in non-literary than in literary translation and when English was the target language.

LLMs, however, possess an advantage that could potentially place them ahead of NMT systems. Through their ability to understand language, incorporate knowledge and adapt to specific context upon interaction with users through prompts, they can be fine-tuned and enhance translation quality. Zeng & Liang (2024) evaluated the translation quality and adaptability of GPT-4 against Google Translate and found that the former outperformed the latter, and even more so after engineering prompts to create a context and offer additional information for the content of the translation.

Obeidat et al. (2024) state that even though Google Translate has established itself on the market as one of the most used translation tools since its launch in 2006 and its accuracy is constantly improved with deep learning models, "it has yet to master cultural nuances and figurative aspects of communication." (Obeidat et al. 2024: 5) They, as well, put it to the test against the two LLMs, ChatGPT and Gemini, when it comes to the translation of idiomatic expressions from English into Arabic. ChatGPT and Gemini proved to be way more flexible. The outcomes were the following: Gemini used the most idiomatic language, ChatGPT led when it came to sense-based translation, and Google Translate scored very high in terms of word-for-word translation.

4.4. Confidentiality Policy

Having discussed the confidentiality obligations of interpreters in their different codes of ethics, this section examines how S2S translation systems address similar concerns. As there is no code of ethics for S2S translation, the General Data Protection Regulation (GDPR) (EU 2016), an EU-wide law that governs how personal data is processed and protected, can be considered a parallel to the interpreters' code of ethics presented in Chapter 2.3.

Medical interpreting inherently involves health data, which means any data relating to the physical or mental health of a person. In the GDPR, health data is classified as a special category of personal data, and its processing is generally prohibited, except for medical diagnosis, the provision of health or social care, or treatment. Access to such data should be permitted only to professionals legally bound by professional secrecy (EU 2016).

As the GDPR is designed to be technologically neutral, it applies the same rules to any data processing agent, whether human or non-human. As a result, the use of a S2S translation tool or a LLM for medical purposes requires high standards of security to protect patient confidentiality, including measures such as anonymisation. Responsibility for ensuring these protections lies with the entity that determines the purpose and method of patient data processing, i.e. the hospital or medical facility (EU 2016).

Another legal document with the purpose of regulating the use of AI is the EU AI Act (EU 2024). The regulation classifies general-purpose AI systems used for translation as having a limited risk factor, seeing that they are meant for general purposes, but fails to account for their use in high-stakes settings such as the medical field.

Despite the existence of these legal frameworks (GDPR, the EU AI Act), there are some ethical considerations regarding the use of S2S translation tools, especially when they run on third-party cloud servers. Moniz & Parra Escartín (2023) highlight some of the potential risks, starting with the fact that speech in itself is a biometric signal that can give away the "preferences, personality traits, mood, health, political opinions, among other data such as gender, age range, height, accent" (Moniz & Parra Escartín 2023: 215) of the user. To prevent the disclosure of such personal information, they state that speech should be legally regarded as "Personally Identifiable Information (PII)" (Moniz & Parra Escartín 2023: 216). Despite S2S technologies being around for a while already, Moniz & Parra Escartín (2023) recognise that most users are oblivious to the potential dangers such as voice cloning, spoofing, incrimination, defamation, misinformation or the embedding of inaudible commands in the speech output.

When it comes to LLMs in particular, Chen et al. (2025) and Van Kolfschooten et al. (2025) mention the same risks of revealing PII, including names, addresses, bank information, sensitive health records, and data stored on third-party cloud servers, where it may be reused for training. Van Kolfschooten et al. (2025) bring up additional ethical concerns, such as the potential of patients giving uninformed consent in case of undetected accuracy errors, and accuracy being linked to biases reflected from the training data.

OpenAI acknowledges that it collects and uses content submitted via its services to improve the model's performance, to "better understand user needs and preferences" (OpenAI 2026). To mitigate security risks, they provide some measures, such as opting out of having their content used for training through a privacy portal or by adjusting the data controls in ChatGPT, using a temporary chat, reducing personal information or refraining from providing feedback, thus placing the responsibility in the hands of the user.

5. Quality Evaluation

This chapter examines the concept of quality in translation, with a particular focus on interpreting. It introduces frameworks and metrics used in translation and interpreting studies, emphasising both linguistic accuracy and communicative effectiveness.

Assessing translation and interpreting quality is a multifaceted process, as there are multiple parties, including translators and interpreters themselves, who form perceptions about the act. However, regardless of the setting and mode, "ideational clarity, linguistic acceptability and terminological accuracy as well as fidelity on one side, and appropriate professional behaviour on the other" (Gile 2009: 39) are general criteria that indicate high quality in translation.

As in translation, quality analysis in interpreting focuses on the degree of correspondence between the original statements and the interpretations regarding content, vocabulary, stylistic devices, syntax, grammatical aspects, etc. Research on the quality assessment of interpreting started only around the 1980s and focused on conference interpreting, both from the perspective of interpreters and listeners (Pöchhacker 2001: 411). An early model introduced by Pöchhacker (2001) comprised the lexico-semantic as well as the socio-pragmatic dimension of interpreting. He suggested to focus on accuracy, clarity and fidelity as product-oriented parameters, and linguistic acceptability and stylistic correctness as listener-oriented ones, but even more so, on the success of the communicative event in itself.

While quality assessment in MT has benefited from well-established and systematically developed frameworks, and research in conference interpreting has increasingly addressed issues of quality assessment, community interpreting remains comparatively underexplored. To date, no widely accepted framework for quality assessment in community interpreting has been established (Hofer-Falk 2023: 62). According to Guo et al. (2024), little research has focused on the listener's perception of quality rather than on the quality of the interpreted output itself. Some studies found that features such as non-native accents, monotonous intonation, erratic prosody, silent pauses, and hesitation (Guo et al. 2024: 2) can lower the perceived quality of an interpretation for a listener. Since the nature of consecutive interpreting entails a more direct contact between the interpreter and the listeners than simultaneous interpreting, which automatically influences the perception of quality. Guo et al. (2024) considered this aspect and proposed a framework for analysing the listener's quality perception in consecutive interpreting scenarios, centred around the following six variables: "listeners' quality expectations, experiences with CI [Consecutive Interpreting], domain knowledge, perceived dependence on

CI, perceived characteristics of the interpreter, and perceived communicative effect" (Guo et al. 2024: 2).

When it comes to medical interpreting, the most important goal is to achieve positive patient outcomes. Hamai et al. (2024) noted that interpreters possess the ability to enhance the relationship and communication between healthcare providers and patients, and facilitate patient advocacy and understanding. Starting from the assumption that, by making active decisions, interpreters can positively or negatively influence the source utterance, which can directly impact the quality of interpreting, Hamai et al. (2024) proposed a framework to assess the quality of interpreting in this specific context, accounting for the interactional aspect of interpreting and the missing representation of the positive effect that proactive interpreters can have in such situations (see Fig. 5). They viewed the interpreted output as altered when the meaning of the input was changed, or unaltered when the segment was interpreted accurately. In an unaltered segment, the interpretation closely resembles the original message, expressing the same ideas, maintaining the full intent of the content and conveying the underlying medical concept using the correct medical terminology (Hamai et al. 2024: 2). The altered segments were divided into "clinically negative or positive and classified into four types: omission, addition, substitution, or voluntary intervention" (Hamai et al. 2024: 3). Positive adjustments refer to instances where the interpreter actively supports accuracy by making corrections, offering clarifications, easing communication or bridging cultural gaps. They also enhance patient comprehension and reassurance through simplification, paraphrasing, or clearer explanation. In addition, they strengthen the patient-physician relationship by fostering empathy, comfort, and a less harsh tone (Hamai et al. 2024: 2).

Negative alterations were further split depending on their potentially clinically significant or insignificant character. Negative but insignificant instances involve minor changes such as adding, omitting, or substituting words or phrases, as well as subjective interpreting, false fluency, language mistakes or the loss of cultural nuance. They may also entail partial misunderstandings by either party, caused by mistranslation or incorrect medical terminology, as well as role changes, or other errors that do not influence the message intent (Hamai et al. 2024: 2). The potentially clinically significant alterations are errors that can impact medical care, patient understanding and comfort or the patient-physician relationship. These include errors referring to biomedical aspects, such as misdiagnosis, delayed diagnosis, unnecessary examinations, impaired clinical decision-making, inaccurate recording of present illness or medical history, and misrepresentation of diagnostic or therapeutic interventions, which can cause uncertainty regarding the severity of the clinical outcome. The interaction

between patient and physician can suffer if empathy is reduced, collaboration is disrupted, the clinician's authority is diminished, or a harsher tone is conveyed.

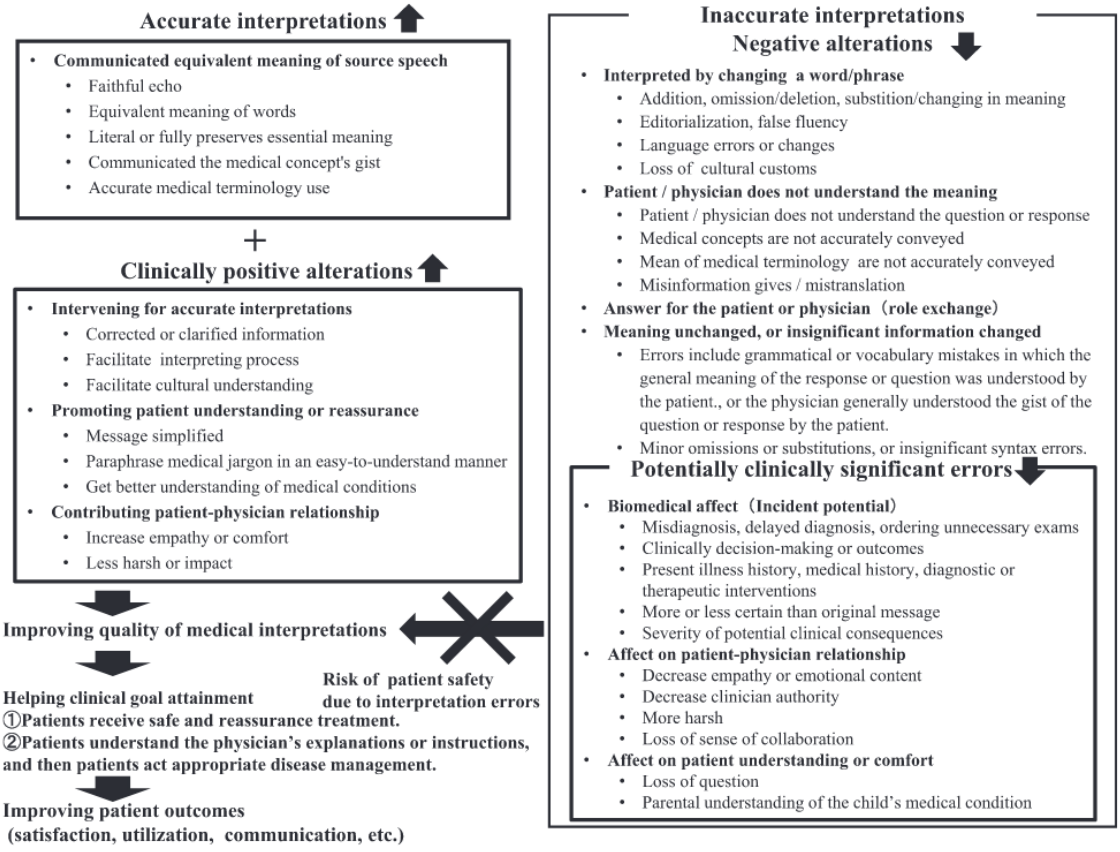


Figure 5. Medical Interpreting Quality Assessment Framework proposed by Hamai et al. (2024: 2)

With the goal of developing a standardised quality framework, fit for evaluation of human as well as machine translations, the European Union funded the QTLaunchPad project, which resulted in the creation of the Multidimensional Quality Metrics (MQM) system (Lommel et al. 2014). According to Lommel et al. (2014), MQM is a rather generalised framework compared to its predecessors, but it allows for customisation and comparison between individual metrics. It is constructed based on hierarchical lists of issue types and can be applied to any language pair. In total, the 2014 framework comprised 114 issue types, organised in main nodes which are Accuracy, Fluency, Verity, Design, Internationalisation, Compatibility and Other, as a customisable category for additional types of issues. Each of these top-level categories has its own set of child issues; for example, in the case of Accuracy, there is Mistranslation, Omission, Untranslated and Addition (Lommel 2014). Furthermore, the translation errors are also divided into three severity levels: minor, major, and critical.

After being published, the framework became widely popular and established itself as a way of evaluating MT and comparing it to human translation (Lommel et al. 2024). It

remained as relevant even after the emergence of NMT, and ten years after the publication of the 2014 framework, Lommel et al. (2024) came up with an updated version that can even serve as a quality evaluation framework for AI-generated translations. While MQM 1.0 only allowed for a rather limited scoring model calculated as a proportion of the sample size, MQM 2.0 introduced two new scoring models that take text length into account and represent the quality more accurately.

For MQM 2.0, Lommel et al. (2024) discern between MQM Core, evolved from a set of the original subtypes of issues that was widely used and called the DQF subset, and MQM Full, comprising all issue types, including some that have been added to address recent technical developments such as "Misrepresentation of technical relationship, False friend, MT hallucination" (Lommel et al. 2024: 5) as subtypes of Mistranslation.

6. Related Work

The following section presents some of the existing literature relevant to the topics of this thesis. It provides an overview of previous research on the use of translation tools in the medical context, highlighting the main methodological approaches and findings within the field and serves to identify the research gaps addressed in this thesis.

Vieira et al. (2021) conducted a literature review, with the aim of providing the first clear overview of the risks of unregulated use of MT in relation to medical and legal communication. The goal was to improve the understanding of MT's risks in these fields and stimulate cross-disciplinary research on the societal impacts of using such technologies. For this, official documents and published research on the use of MT in healthcare and legal settings have been analysed, resulting in a qualitative meta-analysis of the existing literature. The main aspects considered in the study were the perception, usage and effects of MT in legal and medical settings. Vieira et al. (2021) found that use recommendations for MT in healthcare vary inconsistently from country to country, showing a lack of consensus regarding potential health risks. Another problematic aspect of MT in the medical field presented in the study is that some language pairs (especially English and Western European languages) show higher accuracy than other less spoken languages, thus putting the patients who speak less-known languages at a greater risk. As for the legal field, Vieira et al. (2021) identified that many law professionals lack awareness when it comes to MT risks regarding confidentiality, and that, despite the serious implications, there are few official recommendations. Lastly, Vieira et al. (2021) noted that research on MT use in healthcare and law has been mainly conducted by the medical and legal communities, and thus, fails to encompass the dimension of language translation. This all highlights the need for interdisciplinary research in order to raise awareness in society regarding the advantages and limitations of MT and the aggravating social inequalities that the current use is causing (Vieira et al. 2021).

Given that in the medical context, accurate rendition is crucial to prevent mistakes and save lives, Khoong et al. (2019) attempted to assess the use of GT to translate emergency department discharge instructions from English into Spanish and Chinese by comparing the original English versions to back-translations done by professional translators starting from the Spanish and Chinese MT output. Their corpus consisted of 100 free-text emergency department discharge instructions, with a special focus on medication changes and common complaints. Each sentence in those texts was analysed according to content, readability and medical-specific

terms. The texts were then translated into Spanish and Chinese using GT, and back into English by human translators in order to assess the accuracy and categorise the errors (Khoong et al. 2019). For accuracy, the overall meaning was considered and evaluated by both clinicians and translators using a binary system of accurate-inaccurate, and a similar binary system was used to categorise the potential harm of the inaccurate translations as "clinically significant/life-threatening vs clinically nonsignificant/no harm" (Khoong et al. 2019: 1), done only by clinicians. The results showed that the translation between Spanish and English showed a higher accuracy (92%) and a smaller number of mistakes that can cause clinical harm (2%) than those between English and Chinese, with 81% accuracy and 8% clinical harm potential (Khoong et al. 2019). Factors that influenced the translation accuracy were errors in the original English text, and the use of an abundance of medical-specific terms increased the potential of clinically significant mistakes (Khoong et al. 2019).

By 2016, GT was already considered an alternative to human interpreting services in the medical field (Leite et al. 2016: 967). In order to discuss the validity and usability of this tool in medical services, Leite et al. (2016) summarised a case report in which GT was used to communicate with a patient in the psychiatric unit. The patient was a Ukrainian male, admitted to Hospital Magalhães Lemos in Porto, Portugal, after he displayed aggressive behaviour and claimed to see and hear his grandmother, who had passed away. The patient had been living in Portugal for only 5 months at that time and did not speak Portuguese. Because of financial reasons, interpreting services were not available. Initially, the clinical team tried to communicate with the patient through his mother, who acted as an interpreter, but the patient did not trust his mother's intentions, so the clinical team ended up using GT to communicate with him. This language mediation tool, which was impartial in the eyes of the patient, and with which he was willing to cooperate, enabled the healthcare providers to complete a medical examination, treat the patient in 15 days and successfully discharge him with a prescription of antipsychotic medication (Leite et al. 2016). Even though this is a story of success, the report concludes by stating that the only validated method of interpreting in a medical setting is via a human interpreter, and tools like GT should be the last resort (Leite et al. 2016).

Moving on to S2S translation technology being used in medical contexts, Hudelson and Chappuis (2024) conducted one of the few experiments in a real and natural setting. In hopes of determining which factors can facilitate or hinder communication via S2S translation technology between patients and healthcare providers, they attended 60 consultations where either the Microsoft Translator App (Microsoft Corporation 2025) or a S2S translation device called Pocketalk (Sourcenext Corporation 2025) were used in 18 different languages at the

Geneva University Hospitals. They tried to determine if the consultation goal had been met by evaluating the "understanding and satisfaction, willingness to use MT again, difficulties encountered; factors affecting communication when using MT." (Hudelson & Chappuis 2024: 1095). As a result, the consultations were successful in 82.7% of the cases, but only 53.8% were satisfied with the communication method. Most problems were posed by the speech recognition, in cases where the speech was not fluent, where there was too much emphasis on intonation for granting meaning, when numbers were enumerated, or when the speaker had a strong dialect or used words in other languages. Another issue was posed by the additional task of remembering to turn on the microphone of the app at the right time (Hudelson & Chappuis 2024).

Overall, according to Hudelson and Chappuis (2024), 86% of patients declared that the communication via Microsoft Translator (Microsoft Corporation 2025) or Pocketalk (Sourcenext Corporation 2025) was easy, and most participants were willing to use the tools in the future. Participants preferred this option over telephone interpreting and noted that being able to read the translation and not having to depend on another person represented advantages. On the other hand, using the tools required them to change the way they would normally speak, and the novelty of the tools and getting accustomed to them could lead to longer consultation durations.

In order to achieve the consultation goal when there is a language barrier between the patient and the healthcare provider, the medical staff need to know how to act in a triadic-type setting. As this technical aid for communicating with patients does not only concern the translation community, as mentioned at the beginning of this chapter, efforts have also been made in medical research. During a patient communication class, Hermann-Werner et al. (2021) conducted a 90-minute simulation of a doctor-patient interaction using the App iTranslate Converse (iTranslate GmbH 2025), to develop a proof-of-concept pilot that assesses its perceived usability, helpfulness, and meaningfulness (Herrmann-Werner et al. 2021).

In class, students were divided into groups, and some had to play a German-speaking physician instructed to find out the "history of present illness (HPI)" (Herrmann-Werner et al. 2021: 3) from an Arabic-speaking patient with back pain and nausea at 3 AM. After the simulation, students were given a questionnaire where they evaluated how helpful, intuitive, informative, accurate, recommendable, and applicable, on a Likert scale ranging from 1 (don't agree at all) to 7 (completely agree); and then they could freely describe some general impressions, benefits, risks and suggestions for the translation app (Herrmann-Werner et al. 2021). Out of 111 students, 76 participated in the questionnaire and assessed the helpfulness,

recommendability and applicability of the app as average, the accuracy as below average and the intuitiveness and informativeness as above average. In addition, translation errors, lack of empathy, data protection and technical reliability represented concerns of the students, while the benefits were cost-effectiveness and usefulness. As a suggestion, training the tool with medical-specific vocabulary and adding an image visualisation function was considered to improve the quality of the tool (Herrmann-Werner et al. 2021).

One of the most recent studies published on the use of GT in a medical context is Brandenberger et al. (2025), which assessed the accuracy of GT and the legal and policy implications. This was examined based on a simulated paediatric emergency scenario in which a non-English speaking caregiver of a child with respiratory illness and fever had to communicate with an English-speaking healthcare provider. The caregiver was played by a professional interpreter, and the physician by a bilingual healthcare provider, fluent in English and one of the tested languages: Spanish, French, Urdu, Arabic, and Mandarin.

Using the newest version 6.55.0 of GT in conversation mode, the parties could speak into the microphone and then check the transcription of the utterance before it was translated into the target language. This allowed for corrections and increased the quality of translation. When evaluating accuracy, the parameters taken into account were whether the segment was correctly interpreted from the first try, after one repetition, misinterpreted or if it was an undetected misinterpretation. From there on, the segments that were correctly interpreted on the first try or after correction were categorised as accurate (Brandenberger et al. 2025).

As a result, GT showed a strong translation performance, reaching an accuracy of 95.4% in French, 91.7% in Arabic, 88.1% in Mandarin, 86.2% in Spanish and 83.5% in Urdu (Brandenberger et al. 2025). The challenges were posed by a strong dialect, and recurring mistakes were pronoun misinterpretations.

As for the legal concerns, Brandenberger et al. (2025) identified the lack of concrete regulations in the medical field that grant access to interpretation services, as there are for the legal context. Other risks involve data privacy, consent, and malpractice when using AI-based translation tools (Brandenberger et al. 2025).

Brandenberger et al. (2025) conclude that GT could effectively support communication in specific low-acuity paediatric emergency scenarios. However, there is a need for more awareness of the equity considerations and monitoring of unstructured usage of such tools. Before using GT, there needs to be informed consent, assurance of data anonymity, and the use of secure cloud solutions is recommended to reduce legal and data risks (Brandenberger et al. 2025: 6).

Since the emergence of LLMs, the translation technology landscape has been facing a paradigm shift. In their analysis of this phenomenon, Lyu et al. (2024) discovered that advanced general-purpose LLM models surpassed the document translation performance of specialised MT systems in various domains, are contextually aware, which makes them capable of maintaining coherence in long texts, prove great flexibility in adapting translation to different styles and purposes through prompts and allow for the incorporation of translation memories and multi-modal input. One of the concerns of such technologies lies, however, in their ability to guard privacy (Lyu et al. 2024).

Ishida et al. (2025) managed to show that LLMs are now capable of mastering the challenges that MT has been facing in conversational translation over the past few decades. By conducting a literature review, they found that lexical ambiguity, colloquial language and cultural context were the main obstacles for earlier translation systems, and proposed the solution of the "human interpreter metaphor" (Ishida et al. 2025: 2). By interacting with the user, when a problem or misunderstanding arises, as a human interpreter would do, LLMs could improve translation quality and become translation agents (Ishida et al. 2025: 2). The main findings of this study were that LLMs were capable of disambiguating words through interventions, incorporating glossaries, maintaining terminological consistency, handling slang and colloquialisms, deriving meaning from context and improving translation quality by providing back translation and soliciting clarifications.

7. Methodology

The following chapter presents the aim of this research in accordance with the research questions, the research design and the methodology. As the technological shift influences almost all existing domains, and interpreting is inevitably also subject to it, the objective of this thesis is to examine the perceived understanding and interpreting quality in medical consultations mediated by a human interpreter versus an LLM-based S2S translation system by conducting deductive research (Saldanha 2014) to answer the following research questions:

> RQ1: How does S2S translation via ChatGPT compare to human interpreting when assessing the perceived understanding of healthcare providers and patients during medical consultations in German and Romanian?
>
> RQ2: What differences in quality emerge between human interpreters and S2S translation via ChatGPT when analysed through the framework proposed by Hamai et al. (2024) and MQM?

Additionally, the thesis explores potential ethical considerations, including confidentiality and data privacy, associated with each interpreting mode.

7.1. Research Design

To investigate the posed research questions, four simulations of medical consultations between a German-speaking healthcare provider (A) and two Romanian-speaking Patients (P1 and P2) were organised. Two consultations were mediated by a human interpreter, and two by a S2S translation system. Following the simulations, a group interview was conducted with the patients, and a separate one was carried out with the healthcare provider.

Table 1. Overview of the simulated consultations

Consultation	Duration	Segments	Mediation Mode	Participants
C1	11m 46s	93	Human interpreter (D1.1)	Doctor (A1.1), Patient (P1.1), Interpreter
C2	8m 45s	69	S2S (ChatGPT)	Doctor (A1.2), Patient (P1.2)
C3	9m 18s	117	Human Interpreter (D2.1)	Doctor (A2.1), Patient (P2.1), Interpreter
C4	9m 40s	52	S2S (ChatGPT)	Doctor (A2.2), Patient (P2.2)

7.1.1. Participant Selection and Preparations

Particularly in this field, it is quite challenging to conduct research in natural settings due to privacy concerns and the emotional implications for patients. Thus, studies focusing on S2S technology used in medical care, like Brandenberger et al. (2025) and Herrmann-Werner et al. (2021) focused on scripted interactions in controlled environments (Hudelson & Chappuis 2024). For the same reasons mentioned, this research will also rely on simulations. In order for the experiment to be relevant, the participants had to fulfil certain criteria. For example, the healthcare provider had to be German-speaking and the patients Romanian-speaking, with no or low understanding of Romanian and German, respectively. The human interpreter had to be fluent in both languages, professionally trained, and have had prior interpreting experience in the medical context. For this reason, this thesis employed "*purposeful sampling*" (Whyatt et al. 2025: 22) or "*convenience sampling*" (Whyatt et al. 2025: 22).

For the simulations, all participants were provided ten days ahead of the simulation with the scenarios (see Appendix – Scenarios) they were going to play and with instructions, so that they could become familiar with the script and situation. In their scenarios, the patients were given detailed information about the symptoms they were experiencing over a certain period of time, and their relevant medical history. The scenarios followed the same structure: a brief introduction summarising the case; a context section containing details about how the symptoms developed and possible underlying conditions; and the task of communicating their problem to the healthcare provider and finding out the next steps to the best of their ability. The patients were informed that they could adapt the scenarios as they seemed fit, provided the symptoms stayed consistent, and the scenario remained plausible and coherent. This allowed for a more realistic interaction while maintaining the structure of the conversation required for the analysis.

The healthcare provider also received a short scenario at the same time as the patients did. Their scenario specified that they are a doctor on a shift at the emergency department at various times of the day, that there is an incoming patient who does not speak the same language as them and whether there is a human interpreter available at that moment, or they have to use the ad-hoc solution of a S2S translation system on their phone and which one. The healthcare provider's task was to find out the patient's medical history and communicate the next steps to them within approximately 10 minutes.

Prior to participating in the experiment, all participants were informed about the purpose of the thesis. They received a briefing about their roles, the consultation process, and how the data will be gathered and anonymised.

7.1.2.Speech-to-Speech Translation Tool Choice

A range of translation technologies is now available, encompassing both NMT and large LLM-based systems, such as Google Translate (Google 2025b), DeepL (DeepL SE 2025), Microsoft Translator (Microsoft Corporation 2025), and Gemini (Google 2025a). According to Hidayati & Nihayah (2024), Li (2024), Zeng & Liang (2024), Obeidat et al. (2024) and Hua et al. (2024), ChatGPT (OpenAI 2025a) is currently perceived by users to be more context-aware, offering more natural, sense-based and coherent translations than traditional tools like Google Translate (Google 2025b), and possesses an interactive ability which significantly improves both accuracy and fluency by incorporating additional knowledge. It offers speech recognition and speech synthesis from and into Romanian in the free version and was, thus, chosen as the perfect candidate for the experimental part of this thesis. As the simulations took place in December 2025, the newest version available of ChatGPT at that time, GPT-5.2, was used (OpenAI 2025b).

7.1.3.Data Collection

The data was collected through non-participant observations and audio-recordings of the simulations, as well as the interviews via the Samsung Voice Recorder application on a Samsung Galaxy Tab 8 tablet and on an iPad Air 6 as a backup. Afterwards, the recordings were transcribed using a combination of the automatic transcription provided by the same application and manual review and correction. In addition, transcripts generated by ChatGPT were included in the analysis to explore the difference between its transcription and translation output.

Following the simulation, the healthcare provider and the patients were interviewed to assess the general impressions regarding quality, understandability and trustworthiness of the communication mediator, providing insight into the perception of the S2S tool during interpreting medical consultations. As semi-structured interviews “follow a flexible guide that allows for dynamic questioning based on the interviewee’s responses” (Dorner et al. 2025: 69), this was the adopted method for the present thesis. The interviews were divided into four sections: General impressions of the communication, comparison between the two

interpretation methods, empathy, quality and error perception and final questions. The interview guide (see Appendix – Interview Questions) provided the backbone of the interviews; however, the questions were adapted or expanded upon in accordance with the patients' input.

The interview starts with a general introduction that assesses prior experience with human interpreters or S2S translation systems and the overall communication experience during the consultations. Seeing as there is no feasible way of measuring the concrete understanding of the participants, the focus of the interview lies on gathering a general feeling of understanding. A smooth conversation flow, accurate responses and questions, and the self-appreciation of the participants as to whether they felt understood by the other party will be indicators of the degree of perceived understandability. The most important aspect of a medical consultation is the health of the patient, so the aim is to also get an impression about the confidence of the patient in following the doctor's instructions, and the certainty and satisfaction of accomplishing the task for the healthcare provider. Empathy is a subtopic which will also be addressed in the questions for the participants, as it is an important factor for patient outcomes. In the last section, the emphasis lies on the willingness of both to use either of the services. The participants were asked to motivate some of their answers.

As the interview was semi-structured, the questions developed spontaneously but followed the aforementioned framework. For the evaluation and validation of the simulation and the prepared interview questions, a pilot study was conducted beforehand.

7.1.4.Data Analysis

This study adopts a mixed-methods approach, combining a quantitative error and alteration annotation with a qualitative interview discussion. To address the two research questions of this thesis, the quality of both the human interpretation and S2S translation was assessed by annotating the alterations in the transcripts of the recorded conversations using the framework for assessing medical interpreting quality proposed by Hamai et al. (2024) as well as the errors according to the MQM framework, as described by Lommel et al. (2024).

Since MQM is a framework created for translation quality assessment, not all the issue types are applicable to an interpreting setting. Given that this thesis focuses on community interpreting, the most relevant error types and subtypes that will be investigated are:

Accuracy

- Mistranslation
- Addition
- Omission

Style
- Language register
- Awkward style

Linguistic conventions
- Grammar

Additionally, the interviews with the patients and the healthcare provider following the consultations were used to gather the subjective impressions of the participants regarding the overall quality of the interpretations, but were also focused on exploring their perceived feeling of understanding. These subjective perceptions were associated with the objective evidence of understanding identified in the transcripts of the consultations, according to the typology proposed by Clark & Schaefer (1989). However, because the analysis was based on audio recordings of the simulations, the weakest evidence type, i.e. continued attention, was not considered. The correlation of the error analysis and interview data allowed for triangulation, investigating both research questions through two different methods (Saldanha 2014: 23).

For the annotation, the transcript of each consultation was divided into segments, each segment corresponding to a speaker turn. If overlapping speech occurred, it was marked between (…), indicating which participant was speaking. Corrections or rephrasing were marked with […]. In the scenarios where the S2S translation tool was used, it occurred that the transcript provided by ChatGPT was different from the audio input or output, thus presenting an interesting object of analysis. For this reason, the transcript provided by ChatGPT is marked with *Transcript,* and the actual transcript of what the participant said is marked with *Voice*.

A separate analysis table was created for each consultation (C1, C2, C3, C4). The first column shows the number of the segment, the second contains the transcript, and the third and fourth columns were used for the annotation according to Hamai et al. (2024) and MQM, respectively, and the evidence of understanding by the typology of Clark & Schaefer (1989) is depicted in the last column. Cells for which no analysis was applicable are marked with a dash (-). The following abbreviations were used for the annotation and evidence of understanding:

Table 2. Abbreviations

Alteration Type Hamai et al. (2024)	Abbr.	Issue Type MQM	Abbr.	Evidence Type Clark & Schaefer (1989)	Abbr.
Unaltered/accurate	p	Accuracy - Mistranslation	a.mt	Initiation of the relevant next contribution	IRNC
Addition positive	a+	Accuracy - Addition	a.a	Acknowledgement	ACK
Omission positive	o+	Accuracy - Omission	a.o	Demonstration	DEMO
Substitution positive	s+	Style - Language register	s.lr	Display	DISP
Voluntary intervention positive	v+	Style - Awkward style	s.as		
Addition	a	Language register - Grammar	lr.g		
Omission	o				
Substitution	s				
Voluntary intervention	v				
Addition negative	a-				
Omission negative	o-				
Substitution negative	s-				
Voluntary intervention negative	v-				

The interview uses the same abbreviations for the participants as shown in Table 1, and the letter *I* to represent the interviewer. In the transcript of the consultations, informal language or dialect was transcribed as in the original speech, for it plays a role in the analysis, but in the interview, informal language was corrected to standard German or Romanian. Lastly, the interviews were both translated into English, as it is the language of the present thesis and identifying information, including participants' names were modified to ensure anonymisation.

7.1.5.Scoring Model

The quantitative analysis of this thesis is represented by the quality assessment based on the framework for medical interpreting quality proposed by Hamai et al. (2024) and MQM, and the understanding of evidence identification according to Clark & Schaefer (1989). For both frameworks, only the final output was considered, so corrections are not viewed as alterations or errors. In the following, the three different scoring models used in this thesis will be described.

The first scoring model was based on the one proposed by Hamai et al. (2024). For this, the total number of each alteration type across all the segments of one consultation was added up and divided by the total number of translated segments. To get an idea of which alterations are

most prevalent for each interpreting method, taking the different lengths of the corpora into consideration, a score normalised to 100 segments was calculated for each alteration subtype using the following formula:

$$\text{Score}_{\text{alteration}} = \frac{\text{Segments}_{\text{alteration}}}{\text{Total Translated Segments}} \text{ x } 100$$

The sum of all scores of the positive alteration scores in each consultation represents the final score of positive alterations, whereas the sum of all negative (clinically significant and insignificant) alterations in each consultation determines the final score of negative alterations.

The scoring model for the MQM framework was adapted from Lommel et al. (2024) to fit the model to community interpreting, rather than translation. The error types were identified at a segment level; each type was counted only once per segment, if present. Instead of averaging the errors over the word count, the errors were averaged over the segment count. This score was normalised to 100 segments as well, to facilitate the comparison to the first framework. The identified errors were annotated with both the MQM issue subtype and a severity level of either minor, major or critical. The severity levels were assigned according to the weighted scoring model based on Lommel et al. (2024) as:

- Minor (weight = 1): Errors that do not significantly affect the meaning, usability, understandability or reliability of the content for its intended purpose, but deviate from expected linguistic and stylistic norms.
- Major (weight = 5): Errors that considerably impair the meaning, usability, understandability or reliability of the content for its intended purpose.
- Critical (weight = 10): Errors that completely misrepresent the content for its intended purpose and lead to grave misunderstandings that pose a serious health risk.

Because harmful errors were considered to contribute proportionally more to the overall quality assessment, each annotated error was multiplied by its corresponding severity weight, which resulted in the following formula for the second scoring model:

$$\text{Score}_{\text{error}} = \frac{\text{Segments}_{\text{error}} \text{ x } \text{Weight}_{\text{error}}}{\text{Total Translated Segments}} \text{ x } 100$$

Lastly, the final MQM score was calculated for each consultation by summing up all the scores for each error type.

To investigate the evidence of understanding, the total number of each evidence type in all the segments per consultation was summed up and divided by the total number of segments, then

multiplied by 100 to reach an averaged normalised score per 100 segments using the following formula:

$$\text{Score}_{\text{evidence}} = \frac{\text{Segments}_{\text{evidence}}}{\text{Total Segments}} \text{ x } 100$$

8. Research Results and Discussion

This chapter presents the findings of this thesis based on the analysis table of the four simulated medical consultations and the subsequent interviews with the participants, followed by a discussion of the results. The next sections are structured according to the two main focal points of the present thesis, i.e. (a) assessing the perception of understanding among German-speaking healthcare providers and Romanian-speaking patients and (b) evaluating the quality of consecutive medical interpreting produced by ChatGPT compared to a human interpreter, using an error-based framework (MQM) with an interaction-oriented approach (Hamai et al. (2024)).The consequent discussion will highlight the strengths and limitations of the S2S translation system and reveal the participants' preference between the human interpreter and ChatGPT.

8.1. RQ1: Perception of Understanding

8.1.1.Human Interpreter

During the two simulated consultations interpreted by the human interpreter, the participants displayed almost every type of evidence of understanding (Table 3).

Table 3. Evidence of Understanding in C1 & C3

Type of Evidence	C1	C3
	Score	Score
IRNC	26.88	23.93
ACK	26.88	21.37
DEMO	2.15	0.00
DISP	1.08	2.56

One of the two most preponderant types of evidence of understanding was the initiation of the next relevant contribution, scoring 26.88 and 23.93, respectively. This was most frequently displayed in both cases by the patient as they responded to the doctor's questions, providing an accurate level of contextually relevant information, meaning that they understood what the doctor was expecting from them through the interpreter. Consequently, most of the doctors' turns contained evidence in the form of verbal acknowledgement such as "Aham", "Alles klar" [All right], "Okay", "Gut" [Good], separate or in various combinations (e.g. C1 segments 7, 27, 31 and C3 segments 5, 9, 55). Several times in C1, the acknowledgement of the healthcare provider or the patient occurred simultaneously with the interpreter's turn, such as in segments 14 or 66:

14.	D1.1	Wahrscheinlich eher eine Acht (**A1.1**: Aham) im Moment.
66.	D1.1	Deci (**A1.1**: Ahm, genau), în prima fază am o imagine de ansamblu. În următorul pas asistentele vă vor lua sânge și vă vor pune o branulă. Da? (**P1.1**: Okay; **A1.1**: Ja?) Intravenoasă.

Excerpt from Appendix 1. Examples of acknowledgement C1

In C3, the patient and the interpreter did not overlap. However, for the healthcare provider, it occurred two times, such as in segment 106:

106.	D2.1	Ich hoffe (**A2.1**: Aham) es wird nichts Schlimmes sein.

Excerpt from Appendix 2. Example of acknowledgement C3

The interpreter also voiced acknowledgement during the turns of other speakers on two occasions in C1, such as in segments 5 (twice) and 56, and one time in C3, in segment 49:

5.	P1.1	Bună, în primul rând, mă numesc George, eram cu mașina, mergeam cu prietena, sunt pe drum de trei zile, și într-o excursie prin Europa, și în a patra zi azi dimineață, când am plecat din Budapesta, am băut doar o cafea la granița cu Austria. Și probabil din cauza asta, dar nu sunt sigur de ce, a început să mai ia o durere de burtă super nasoală (**D1.1**: Aham). Progresiv durerea s-a amplificat. Am ajuns în Viena, ne-am plimbat, dar pe cum a trecut timpul, m-a durut burta și mai mult. Și la un moment dat, pur și simplu simțeam că nu mai puteam să merg (**D1.1**: Aham). Și eram mult prea obosit. Am decis să mă duc înapoi la pensiunea unde eram cazați. Și am crezut că o să-mi revin, dar nu mi-am revenit. În continuare mă simt foarte prost. Am avut stări de greață. Am transpirat foarte mult. M-am pus în pat. N-am mai putut sta în picioare. Și acuma am rugat-o pe prietena mea să mă ducă la spital. De asta sunt aici.
56.	P1.1	Nu am măsurat febra (**D1.1**: Okay). Doar am început să transpir foarte mult, dar mi-a fost cald. (**D1.1**: V-a fost cald?)

49.	P2.1	Este gălbuie (**D2.1**: Aham).

Excerpt from Appendix 3. Examples of acknowledgement C1 & C3

In C3, the interpreter started two of their turns (segments 8 and 54) by signalling that they were able to follow the other speaker and render their message:

8.	D2.1	Aham. Ja, seit zwei Tagen (**A2.1**: Aham) erbreche ich regelmäßig und habe Durchfall.
54.	D2.1	Aham. Vor dem Stuhlgang habe ich ganz schlimme Bauchschmerzen.

Excerpt from Appendix 4. Examples of acknowledgement C3

Evidence in the form of demonstration was only found in C1, displayed by the healthcare provider on two occasions, both times combined with acknowledgements. In segment 38, the

interpreter was specifying the location of the abdominal pain “Rechts unten” [bottom-right], and the doctor rephrased the information for themselves as “Rechter Unterbauch” [lower right abdomen]. In segment 5, the patient mentioned that they sweated, and in segment 54, the healthcare provider returns to this with “Schwitzen haben Sie gesagt” [You mentioned sweating], inquiring more details about the symptom.

39.	A1.1	Rechter Unterbauch, okay, gut. Wie schaut’s aus mit dem Stuhlgang? Irgendwelcher in Richtung Durchfall gehabt auch heute oder die letzten Tage?
54.	A1.1	Gut, okay. Schwitzen haben Sie gesagt…Fieber auch gehabt?

Excerpt from Appendix 5. Examples of demonstration C1

Display, as a type of evidence, is only weakly represented in C1 in segment 27, when the interpreter translates “Nein, keine” [No, none], meaning that the patient does not have any chronic diseases, and the healthcare provider answers “Nein, okay” [No, okay], repeating the same words of the interpreter, thus also exhibiting understanding.

27.	A1.1	Nein, okay. Danke. Der Schmerz: Wenn Sie den beschreiben würden, ist es eher ein Stechen, ein Ziehen, ein Brennen, ein wellenartiger Schmerz?

Excerpt from Appendix 6. Example of display C1

In C3, however, display is present on three separate occasions, all during the interpreter’s turn, as the healthcare provider repeats their exact same words, signalling they are following them and understanding the message:

32.	D2.1	Ich hatte noch zu Hause (**A2.1**: Zuhause, okay), ich musste nicht in die Apotheke gehen.
42.	D2.1	Mal am Tag (**A2.1**: 2-3 Mal am Tag).
70.	D2.1	Beide (**A2.1**: Beide).

Excerpt from Appendix 7. Examples of display C3

8.1.2.ChatGPT

In the consultations mediated through the S2S translation function of ChatGPT, the two weakest forms of evidence of understanding prevailed, with demonstration occurring only once in C2.

Table 4. Evidence of Understanding in C2 & C4

Type of Evidence	C2	C4
	Score	Score
IRNC	31.88	28.85
ACK	2.90	11.54
DEMO	1.45	0.00
DISP	0.00	0.00

As in C1 and C3, the initiation of the relevant next contribution scored the highest (31.88 in C2 and 28.85 in C4) in those two cases as well, indicating the participants' awareness and understanding of the discussion. The patient, however, displayed almost twice as much identifiable evidence of understanding as the healthcare provider. As for acknowledgements, only two were made in C2: once by the patient in segment 60, and once by the doctor in segment 67:

60.	P1.2	Transcript - Voice Okay, super.
67.	A1.2	Alles klar, haben Sie jetzt noch Fragen, bevor wir starten?

Excerpt from Appendix 8. Examples of acknowledgement C2

Segment 60 did not get recognised as a separate turn by ChatGPT and was instead transcribed in segment 61, as part of the doctor's turn, thus resulting in a minor omission. Given that the patient's answer, "Okay, super", was intelligible in all the languages known to the healthcare provider, the acknowledgement reached the target audience and the presentation made by the patient was silently accepted.

C4 contained six acknowledgements, i.e. three from each speaker. The patient displayed acknowledgement in the form of thanking the healthcare provider in the segments 39, 43 and 51, all of them being correctly transmitted. In segments 17 and 37, the acknowledgements of the doctor were accurately rendered by ChatGPT, while in segment 28, the translation tool was interrupted by the premature acknowledgement, thereby eliminating the need for the full translation of the source utterance, as the message was already understood.

28.	ChatGPT	Transcript Nein, ich habe keine Allergien. Voice (**A2.2**: Okay) Nein, ich hab…

Excerpt from Appendix 9. Example of acknowledgement C4

The only stronger form of evidence found in those two consultations is one demonstration from the doctor in segment 12 of C2, when rephrasing the information they got from the patient, that they had a productive cough as “Und der Husten war mit Auswurf” [And the cough was productive].

12.	A1.2	Und der Husten war mit Auswurf. Wie schaut denn das aus, wenn Sie aushusten müssen? Welche Farbe hat denn das?

Excerpt from Appendix 10. Example of demonstration C2

8.2. RQ2: Quality

8.2.1.Human Interpreter

Table 5. Quality of C1 & C3 through Hamai et al. (2024)

Clinical relevance	Error types	C1	C3
Unaltered (accurate)	p	76.60	76.67
Positive			
	a+	0.00	0.00
	o+	4.26	0.00
	s+	14.89	6.67
	v+	2.13	6.67
Negative - Not clinically significant			
	a	0.00	0.00
	o	0.00	3.33
	s	2.13	3.33
	v	0.00	0.00
Negative - Potentially clinically significant			
	a-	0.00	0.00
	o-	0.00	0.00
	s-	2.13	1.67
	v-	0.00	0.00

The two human-interpreted scenarios got scores of 76.60 and 76.67 when it comes to unaltered segments, meaning that about three-quarters of the interpreted segments were accurate representations of the original utterance. As this framework is based on the idea that the

interpreter is an active participant, capable not only of adequately rendering messages but also of improving the overall quality of the conversation, it also takes into consideration the positive alterations of the interpreter. C1 and C3 contain ten and eight segments, respectively, with positive alterations.

Starting with positive omissions, in C1, the interpreter made the decision to eliminate the doctor's self-talk and acknowledgement of understanding from segment 39 and only inquired about the patient's bowel movement, conveying the necessary medical information solely. In segment 92, the interpreter avoided redundancy by only stating "Bun, începem" [Good, we start], because the topic of the upcoming examinations started five turns before, and it was already implied that the doctors would start the examinations. In both cases, the meaning of the utterances was accurately conveyed, without complicating the conversation unnecessarily.

39.	A1.1	Rechter Unterbauch, okay, gut. Wie schaut's aus mit dem Stuhlgang? Irgendwelcher in Richtung Durchfall gehabt auch heute oder die letzten Tage?
40.	D1.1	Cum e cu scaunul? Ați avut diaree astăzi sau în ultimele zile?
91.	A1.1	Gut. Dann fang'ma an jetzt mit den Untersuchungen.
92.	D1.1	Bun, începem.

Excerpt from Appendix 11. Examples of positive omission C1

In segment 59 of C1, the interpreter rephrased the question of the healthcare provider, "Schüttelfrost?" [chills?] formulating a full sentence "Ați avut și frisoare" [Did you also have chills?], in order for the patient to clearly understand the question referring to additional symptoms and for the dialogue to flow naturally.

59.	D1.1	Ați avut și frisoane?

Excerpt from Appendix 12. Example of positive addition C1

The two positive substitutions in segments 16 and 32 from C1 were the active decisions of the interpreter to remedy the confusion of the healthcare provider as to whom they should address in a triadic constellation, changing the third person singular to the second person singular:

31.	A1.1	Aham. Hat er erbrochen auch?
32.	D1.1	Ați și vomitat?

Excerpt from Appendix 13. Example of positive substitution C1

Positive substitutions occurred in C1 in segments 46 and 52 as well. In the first example, the interpreter added the personal remark "über den ich gesagt hab" [the one I mentioned] when telling the doctor that the patient only drank a coffee. The information had indeed been mentioned before by the patient, but in segment 45, they did not explicitly say that they

mentioned the coffee before; this was thus a rephrasing from the interpreter, meant to reference something previously discussed to ease the recollection of information. In segment 52, when the interpreter rendered the inquiry about the health of the patient's friend "Gut, und die Freundin ist gesund?" [Okay, and the friend is healthy?] by adding another question "Prietena e sănătoasă? Ei îi merge bine?" [Is your friend healthy? Is she doing well?], the interpreter did not distort the meaning of the original utterance, but rephrased to establish clarity and a more natural conversation.

On other occasions, positive substitutions served to rephrase longer passages, such as the background story of how the patient started having the symptoms (segment 6) or the doctor's explanation of the next steps (segment 68).

The positive voluntary interventions represent the proactive involvement of the interpreter in the course of the conversation to clarify details that may cause potential misunderstandings, ask for repetition or additional information. In segment 56, the interpreter intervened during the turn of the patient and asked them a clarifying question ("V-a fost cald?" [Were you hot?]) about the last part of their sentence, in order not to miss any important details that might have been overheard or misunderstood:

56.	P1.1	Nu am măsurat febra (**D1.1**: Okay). Doar am început să transpir foarte mult, dar mi-a fost cald. (**D1.1**: V-a fost cald?)

Excerpt from Appendix 14. Example of positive voluntary intervention C1

C3 counted four positive voluntary interventions. In this consultation, there were multiple instances when the interpreter became an active part of the conversation to clarify misunderstandings. In segment 33, the doctor asked about the patient's drinking habits in the last few days, upon which the interpreter intervened in segment 34 and asked if the doctor wanted to know about the alcohol or just hydration. After getting the confirmation that hydration was meant, the interpreter summarised the last turns into a clear question for the patient about how much they had been hydrating themselves in the last few days, which counts as a positive substitution.

33.	A2.1	Wie schaut's aus mit dem Trinken in den letzten Tagen?
34.	D2.1	Wie meinen Sie? (**A2.1:** Also haben Sie...) Akohol, oder?
35.	A2.1	Nein, nein, normal Flüssigkeit.
36.	D2.1	Vizavi de cum vă hidratați. Cât beți? În ultimele zile.

Excerpt from Appendix 15. Example of positive voluntary intervention C3

Similar positive substitutions meant to improve the clarity and flow of discussion are represented in segments 52 and 74 of C3. First, the interpreter uses the medical term "defecație"

[defecation] and then adds “În timpul eliberării scaunului” [During stool release] to provide a more common phrasing for the medical term, promoting the understanding. Secondly, in segment 74, the interpreter rephrases the original utterance completely, while keeping the meaning, making it easier for the doctor to follow.

Segment 104 showcases the subjectiveness of this evaluation framework as it could be seen as both a positive and a negative, but a clinically insignificant substitution. The interpreter is rephrasing the content of segment 93, presenting the next steps in a very clear and structured manner for the patient. When it comes to the list of “Keime, Bakterien, Viren” [Germs, bacteria, viruses], the interpreter uses the umbrella term “microorganism” [microorganisms] to describe all of them. This change facilitates understanding, but lowers the degree of specificity. For this thesis, however, this segment counts as a positive substitution.

Another effort to establish clarity was made in segment 66, when the interpreter, registering the confusion of the patient from his facial expression and intonation, asked the healthcare provider directly what they meant in segment 63 by “Gut, dann wegen dem Sprunggelenk, auf welcher Seite ist das?” [Okay, so about the ankle joint, which side is that on?], as it was unclear if they meant on which ankle or on which side of the ankle the swelling was. After resolving the ambiguity, the interpreter inquired about which of the patient’s ankles was swollen, thus facilitating a better understanding.

Segments 94 and 102 also represent clear examples of the positive voluntary interventions and how good turn management of the interpreter can improve the quality of the interaction. In segment 93, the healthcare provider explains what the next steps and examinations would be, mentioning an ultrasound of the bladder in the end, which led to the question if the patient had not urinated at all for two days. Because the interpreter deemed it important for the doctor to receive the answer to the last question quicker, they stated in turn 94 to the patient that they would ask a short question, and after that topic is cleared, they would return to interpret the rest of segment 93. This way, the healthcare provider could get the necessary information sooner and have more time to mentally prepare the next steps, while the rest of segment 93 would be translated in segment 104.

Moving on from the positive changes, C1 and C3 also display some of the negative, but clinically insignificant alterations. C3 shows two omissions and two substitutions, while C1 contains just one substitution. Segment 4 (C3) omitted the greeting from the patient at the beginning, and their self-assessment of their state “Și da, mă simt foarte prost” [And yes, I feel very bad], at the end. This did not have meaningful negative consequences, as it was clear that the patient was not feeling well, but it still represents an omission. The second omission refers

to segment 96 (C3), where the interpreter left out the detail of the patient not having urinated for the last two days, only mentioning that they could not urinate since the morning after the excursion. The specific amount of two days was missing from this segment, but it only counts as a clinically insignificant error, as this information was already mentioned in previous sections, and the doctor was aware of the timeline.

C3 contained a case of a negative clinically insignificant substitution in segment 76. The healthcare provider's question if the ankle swelling has gone down is inverted and becomes "Gleznele sunt umflate?" [Are the ankles swollen?], to which the patient answers "Da" [Yes]. Because of the ambiguous polarity, this counts as a negative substitution. Since, however, the doctor understood the meaning as the patient intended it, stating that they would examine the ankles afterwards, and no confusion resulted from this, it is not labelled as clinically significant. The last clinically insignificant substitution in C3 was a grammar error in segment 72, where the word "scrântit" [sprained] is incorrectly pronounced as "sclintit" when the interpreter inquired about a possible sprained ankle.

In C1, there was a negative, but clinically insignificant, substitution in segment 28, when the interpreter did not exactly describe the different types of pain as the doctor mentioned them, not precisely conveying the meaning of "Ziehen" [pulling] and separating "wellenartiger Schmerz" [wave-like pain] into "crampe" [cramps] and "valuri" [waves]. The sentence kept the structure of four types of pain being listed, but the meaning was slightly modified, without posing a potential clinical danger, as the patient understood those were examples of how pain could feel, meant to help them describe their pain better.

27.	A1.1	Nein, okay. Danke. Der Schmerz: Wenn Sie den beschreiben würden, ist es eher ein Stechen, ein Ziehen, ein Brennen, ein wellenartiger Schmerz?
28.	D1.1	Cum ați descrie durerea? Este o înțepătură? Este mai degrabă usturime? Este crampe sau este în valuri?

Excerpt from Appendix 16. Example of negative clinically insignificant substitution C1

Lastly, the interpreter also made one negative clinically significant substitution in each consultation. In segment 23 of C1, the doctor asked about pre-existing conditions, which were translated as "boli cornice" [chronic diseases], limiting the meaning of the term. In case the patient suffers from a pre-existing condition that is not a chronic disease, they may not mention it when asked this question. This could negatively impact the diagnosis.

23.	A1.1	Und Vorerkrankungen?
24.	D1.1	Aveți cumva alte boli cronice?

Excerpt from Appendix 17. Example of negative clinically significant substitution C1

The negative clinically significant substitution in segment 108 of C3 refers to the interpreter implying that the “analgezic” [pain medication] will help the patient against nausea. The doctor specified that the patient will be getting pain and nausea medication, but acoustically, they did not emphasise the “and”, so it may be a reason for the confusion. Due to this alteration, the patient was not informed correctly about the treatment they would be receiving, which counts as a clinically significant negative alteration.

Table 6. Quality of C1 & C3 through MQM

Issues	Sub-issues	C1			C3		
		Minor	Major	Critical	Minor	Major	Critical
Accuracy							
	Mistranslation	2.13	10.64	0.00	1.67	8.33	0.00
	Addition	0.00	0.00	0.00	0.00	0.00	0.00
	Omission	0.00	0.00	0.00	3.33	0.00	0.00
Style							
	Language register	0.00	0.00	0.00	0.00	0.00	0.00
	Awkward style	0.00	0.00	0.00	1.67	0.00	0.00
Linguistic conventions							
	Grammar	0.00	0.00	0.00	1.67	0.00	0.00

As MQM focuses on errors that hinder the meaning, usability, understandability or reliability (Lommel et al. 2024) of the translated, or, in this case, interpreted content, there are considerably fewer segments to be discussed from this perspective. In C1, there were one minor and one major error identified, whereas in C3, there were five minor and one major error.

In C1, all the segments that counted as negative alterations in the first framework are also deemed as errors by MQM. Segment 28 is qualified as a minor mistranslation, as the examples of the pain types were not accurately interpreted. Nonetheless, this did not affect the comprehension of the participants and is considered only a minor deviation. On the other hand, in segment 24, where the pre-existing conditions were rendered as chronic diseases, the implications for the purpose of the content could be more serious, transforming this into a case of major mistranslation.

The six errors in C3 have not all been identifiable through the first framework. Since Hamai et al. (2024) created the assessment method specifically for interpreting, the most important criterion was that understanding is facilitated. Segment 2 poses no understanding

problems for the participants, but contains a minor style error from the MQM perspective. The phrasing in the question “De ce veniți la noi aici la urgențe?” is somewhat unnatural and becomes a case of minor awkward style error. The word “scrântit” [sprained], being incorrectly pronounced as “sclintit” in segment 72, is a language convention breach in the form of a minor grammatical error.

The minor omissions in segments 4 and 96 were commented above as also being examples of negative, clinically insignificant omissions, as well as the minor mistranslation in segment 76 about the ambiguity regarding the swelling of the ankles, which counted as a negative, clinically insignificant substitution. The more serious error in segment 108 concerning the medication the patient was going to receive is also weighted by MQM with the severity degree major, for it negatively interferes with the purpose of the original utterance.

8.2.2.ChatGPT

Table 7. Quality of C2 & C4 through Hamai et al. (2024)

Clinical relevance	Error types	C2	C4
Unaltered (accurate)	p	58.33	88.46
Positive			
	a+	0.00	0.00
	o+	0.00	0.00
	s+	2.78	0.00
	v+	0.00	0.00
Negative - Not clinically significant			
	a	2.78	0.00
	o	8.33	0.00
	s	8.33	11.54
	v	0.00	0.00
Negative - Potentially clinically significant			
	a-	0.00	0.00
	o-	8.33	0.00
	s-	11.11	0.00
	v-	0.00	0.00

When comparing the quality of the two ChatGPT-interpreted scenarios, the first thing to notice is that C2 scored considerably lower on the framework of Hamai et al. (2024) in terms of unaltered segments. Additionally, as the LLM cannot take voluntary actions, it is almost impossible to identify positive alterations. Yet segment 50 in C2 contains an alteration, which could be marked as a positive substitution done by the machine. The patient was telling the doctor in segment 49 that they had taken paracetamol, Nurofen, and vitamin C. Nurofen is a brand of pain relief medication that has ibuprofen as the active ingredient (Reckitt 2026). As this brand is popular in Romania (Euromonitor International 2025), the fact that ChatGPT transformed Nurofen directly to the well-known active ingredient, ibuprofen, facilitated the understanding of the healthcare provider. Whether this should be considered a voluntary alteration and if ChatGPT is even capable of such actions is a topic for debate, but for this analysis, it is seen as such.

49.	P1.2	Nu, dar știu că am luat paracetamol și Nurofen și vitamina C.
50.	ChatGPT	Nein, aber ich weiß, dass ich Paracetamol und Ibuprofen sowie Vitamin C eingenommen habe.

Excerpt from Appendix 18. Example of positive substitution C2

Moving on to the negative, clinically insignificant issues, there are two examples of addition. The first case is in segment 18, when ChatGPT added “Alles klar” [All right], which was not part of the conversation, at the same time when the patient was communicating their body temperature. The second case is in segment 62, when ChatGPT did not register the speaker change between the patient’s answer “Okay, super” in segment 60 and the doctor’s turn in segment 61, as it was also not necessary for the participants to have a translation for segment 60. The patient expressed their acknowledgement in English, which the doctor understood. This, however, led to the answer of the patient being added to segment 62, as if it were part of the doctor’s turn in segment 61. Because of this situation, segment 60 was flagged with a negative, clinically insignificant omission. The patient’s answer did not get translated, but it had no negative consequences, as it was still intelligible for the doctor. Similar cases of insignificant omissions occurred in segments 7 and 57, where the patient’s evidence of understanding “Da” [Yes] and “Okay, am înțeles” [Okay, I got it] were left out completely, but since they did not hinder the communicative goals, they are considered insignificant.

In segment 28, ChatGPT also omitted to translate the whole content of segment 27, where the patient stated that they did not have any allergies. Since the translation tool made a three-second pause, the doctor thought it would not produce any translation, so they relied on body language to guess the negation in the patient’s answer, even displaying an

acknowledgement of understanding. In the end, ChatGPT started producing the translation, but interrupted it abruptly, probably because of the intervention of the doctor. Since the meaning was still conveyed, this was not considered an actual omission error in this framework.

27.	P2.2	Nu, nu am alergii.
28.	ChatGPT	Transcript Nein, ich habe keine Allergien. Voice (**A2.2:** Okay) Nein, ich hab…

Excerpt from Appendix 19. Example of interrupted translation C4

A recurring shortcoming of ChatGPT was that it did not register proper nouns correctly. This is visible in segment 2 of C2 and segments 2 and 4 of C4. When they introduced themselves, the name of the doctor was misrepresented phonetically in two different ways in the second segments of both consultations. In segment 4, the patient also introduced themselves. The rendered names were still phonetically close to the original ones, but since they were incorrect, those segments were labelled as negative, clinically insignificant substitutions.

In segment 40 of C2 and 8 of C4, ChatGPT changed the person of the speech. In segment 39, the doctor asked the patient to repeat how many years they had been smoking, adding that they thought the tool did not understand "Das hat es, glaube ich, nicht verstanden" [I believe it did not understand that]. This was translated as "cred ca nu am înțeles" [I believe I did not understand] in segment 40, changing the person from 3rd to 1st, making it seem as though the doctor had trouble following the patient, not the machine. As the request for the patient to repeat was more important, and this message was conveyed, this counts as an insignificant substitution. A similar case happened in segment 8, but this time, ChatGPT switched from direct speech "Aș zice" [I would say] to indirect speech "Er sagt, dass [He says that]. This change did not cause any meaning distortion or hinder the purpose of the content, and thus be counts as an insignificant substitution.

Another example of the same alteration type is the change to an informal address in segment 22 of C2, which led to the question of the doctor seeming less polite. Throughout the consultations, all participants addressed each other using formal pronouns, but in this segment, the informal "Poți să repeți" [Can you repeat] in 2nd person singular was used instead of the formal "Puteți să repetați" [Could you repeat] in 2nd person plural.

Negative alterations with potentially clinically significant consequences were only identified in C2, with numerous instances observed. In segments 25, 30, and 62, important questions of the healthcare provider were not registered by ChatGPT the first time, and thus,

not translated, until they were repeated. In segment 25, the doctor wanted to know if there were any pre-existing conditions, information which is of utmost importance for a medical consultation. This was, however, not translated, and the doctor had to pose the question again after seeing that the tool did not react. This can also be visible in the Excerpt from Appendix 20, as ChatGPT did not transcribe anything for this segment.

25.	A1.2	Transcript - Voice Gibt es Vorerkrankungen bei Ihnen?

Excerpt from Appendix 20. Example of negative, clinically significant omission C2

On the other hand, for the other two negative significant omissions, transcripts were available, and only the audio output was missing. In segment 30, the doctor asked the patient how many cigarettes they smoked per day, but the question was completely left out. The same happened in segment 62, when the information that the patient will be treated accordingly after the examinations and the part where the doctor inquired if the patient had a pulmonologist, did not get translated. Because information that is crucial for clinical understanding and decision-making has been left out, these errors are seen as having potentially clinical significance.

29.	A1.2	Transcript Wie viele Zigaretten rauchen Sie denn am Tag? Voice Wie viele rauchen Sie denn am Tag? Wie viele Zigaretten?
30.	A1.2	Wie viele Zigaretten pro Tag rauchen Sie?
61.	A1.2	Transcript Okay, super. Und je nachdem, was da herauskommt, kriegen Sie dann die entsprechende Therapie. Und eine letzte Frage habe ich noch, bevor wir starten. Haben Sie einen Lungenfacharzt? Voice Und je nachdem, was da herauskommt, kriegen Sie dann die entsprechende Therapie. Und eine letzte Frage habe ich noch, bevor wir starten. Haben Sie einen Lungenfacharzt?
62.	ChatGPT	Transcript Bine, minunat! Și, în funcție de ceea ce va rezulta, veți primi apoi terapia corespunzătoare, iar o ultimă întrebare înainte să începem: Voice Bine, minunat! Și, în funcție de ceea ce (**A1.2**: Ha…)

Excerpt from Appendix 21. Examples of negative, clinically significant omission C2

Segments 20 and 24 contain one of the gravest errors in those consultations, which has been repeated twice. It seems that ChatGPT performs poorly with numbers, as it failed to understand the patient's body temperature of 38.6°C, although they communicated it multiple times,

changing it to 63.6°C. It is noticeable in segment 19 that ChatGPT had trouble understanding the number when the patient was speaking more informally, and transcribed the source utterance in another language than Romanian. Then, in segment 20, it stated a completely unrealistic temperature. Even though the doctor attempted to clarify the misunderstanding with a second question, and the transcript of the segment 23 seems correct, the audio output still translated to 63.6°C, propagating the mistake. For this reason, the segments are labelled as negative, clinically significant substitutions according to Hamai et al. (2024).

19.	P1.2	Transcript Trej šopku šase. Voice 38 cu 6.
20.	ChatGPT	Transcript 36,6 Grad. Voice 63,6 Grad.
23.	P1.2	Transcript 38. 6 grade Celsius. Voice 38 virgulă 6 grade Celsius.
24.	ChatGPT	Transcript 36,6 Grad Celsius. Voice 63 Komma 6 Grad Celsius.

Excerpt from Appendix 22. Examples of negative, clinically significant substitution C2

Another instance of ChatGPT having difficulties recognising numbers was in segment 38. The patient stated, again in a more informal language, that they had been smoking for 14 years, and the transcript of ChatGPT shows nonsensical content, “Depeche Pane”, indicating that it did not comprehend the number. This was translated as “Zum Beispiel seit…” [For example, since…], which confused the doctor.

37.	P1.2	Transcript Depeche Pane. Voice De paișpe ani.
38.	ChatGPT	Zum Beispiel seit…

Excerpt from Appendix 23. Example of negative, clinically significant substitution C2

Upon repeating the number in a formal, clear manner in segment 41, ChatGPT understood the number 14, but translated another part of the source utterance incorrectly in segment 42, causing another clinically significant negative alteration. In addition to the 14 years, the patient also

reported that they do not smoke a lot, but have been doing so for 14 years. The translation of this segment shifts the meaning, conveying that they did not start that long ago.

41.	P1.2	Transcript De 14 ani, cum es a început, nu mult, dar de 14 ani in total. Voice De 14 ani, cum am zis. La început, nu mult, dar de 14 ani în total.
42.	ChatGPT	Seit 14 Jahren rauche ich, es hat nicht vor langer Zeit begonnen, aber seit 14 Jahren insgesamt.

Excerpt from Appendix 24. Example of negative, clinically significant substitution C2

When evaluated through the MQM framework, the ChatGPT-mediated consultations show mostly mistranslations, omissions, and an awkward style, but also some isolated cases of addition and wrong language register. Additionally, in C2, all severity degrees were represented, while in C4, there were exclusively minor errors.

Table 8. Quality of C2 & C4 through MQM

Issues	Sub-issues	C2			C4		
		Minor	Major	Critical	Minor	Major	Critical
Accuracy							
	Mistranslation	5.56	27.78	55.56	7.69	0.00	0.00
	Addition	5.56	0.00	0.00	0.00	0.00	0.00
	Omission	8.33	41.67	0.00	3.85	0.00	0.00
Style							
	Language register	2.78	0.00	0.00	0.00	0.00	0.00
	Awkward style	11.11	0.00	0.00	7.69	0.00	0.00
Linguistic conventions							
	Grammar	0.00	0.00	0.00	0.00	0.00	0.00

The minor mistranslation cases were previously discussed as substitutions with no clinical significance, and are present in segments 2 and 40 of C2 and segments 2 and 4 of C4. In all of those cases, the content of the source utterance was modified, but it had no impact on the purpose of the communication or any clinical detriments.

Segments 38 and 42 in C2, in which the patient was discussing for how long and how much they had been smoking, represent major mistranslation errors when viewed through the MQM framework. These errors considerably change the meaning and the purpose of the source utterance and could pose potential clinical harm. However, since the misunderstanding in

segment 38 was cleared up through additional questions and segment 42 still carries the most important message across, they are labelled in this category.

An even graver error is represented in segments 20 and 24, which have also been previously discussed as negative, clinically significant alterations. Rendering a wrong body temperature during a medical consultation is a case of critical mistranslation, as this completely misrepresents the content and poses a serious health risk.

Segment 18 contains a minor addition, for ChatGPT adds the comment "Alles klar" [All right], which was not present in the source utterance, overlapping the turn of the patient. Segment 62 provides, at the same time, an example of minor addition and major omission. The addition consists of the incorporation of the patient's response "Okay, super" in the translation of the doctor's answer, even though the doctor's answer did not contain these words. The omission in the same segment consists of the explanation that the treatment will be decided after the investigations, and the important question if the patient had a pulmonologist were completely left out. Other major omissions have been identified in segments 25 and 30. Both cases have been discussed above as negative, clinically significant omissions. First, the healthcare provider's question concerning the pre-existing conditions was omitted by ChatGPT, and then the one about the number of cigarettes that the patient was smoking.

Similarly to the first framework, MQM does not evaluate the omissions in segments 7, 57 and 60 as grave, labelling them as minor. The intention of the source utterances was to acknowledge agreement and understanding, and as the patient used English or words of affirmation, they were understood by the healthcare professional, even though they were in Romanian and remained untranslated. For this reason, these errors are categorised as minor.

The MQM framework has a distinct category for style issues, specifically for language register problems. The change from a formal address to an informal one in segment 22 (C2) is a representation of a minor language register error, as the politeness level which characterised all of the consultations is missing.

The last type of minor error that will be discussed is not represented in the framework of Hamai et al. (2024), and is concerning an awkward or unnatural style. Generally, in C2, the Romanian output of ChatGPT was characterised by an unnatural German accent. This was most evident in segments 6 and 36, which led to them being flagged with this type of error. Segments 56 and 59 sounded awkward because the doctor's attempts to establish rapport and understanding through repeating the question "Ja?" [Yes] at the end of multiple sentences were translated word for word, and placed very unnaturally in the sentences in Romanian.

58.	A1.2	Transcript Okay, haben wir alles.Und dann schicken wir sie zum Röntgen, um zu schauen, was in der Lunge los ist. Das könnte zum Beispiel eine Lungenentzündung sein. Und ich werde sie abhören und generell untersuchen. Voice Und dann schicken wir sie zum Röntgen, ja? Um zu schauen, was in der Lunge los ist. Das könnte zum Beispiel eine Lungenentzündung sein, ja? Und ich werde sie abhören und generell untersuchen.
59.	ChatGPT	Și apoi vă vom trimite la radiografie, da, pentru a vedea ce se întâmplă în plămâni, ar putea fi, de exemplu, o pneumonie, da, și vă voi asculta cu stetoscopul și vă voi examina în general.

Excerpt from Appendix 25. Example of awkward style C2

As for C4, segment 28 is categorised as an awkward style error because ChatGPT stopped in the middle of the sentence and did not finish translating it. The reason for this is probably the early overlapping intervention of the healthcare provider. The main message was, however, not affected. The last example is in segment 38, when ChatGPT produced a grammatically correct output, which is, however, not frequently used in natural Romanian speech “sună puțin a unei pancreatite” [sounds like a pancreatitis]. Even though it sounded unnatural for the patient, the content was not altered and remained intelligible.

8.3. Interview Insights

As mentioned in Chapter 6.2.3, the chosen method of collecting the qualitative data was through interviews with the participants. For this, a group interview was carried out with the two participants and a separate one was held with the healthcare provider after the simulations. At the beginning of both interviews, the participants were asked if they had any prior experience with human interpreters or with ChatGPT's interpreting function. Both patients said they had not previously used any of the methods, while the healthcare professional mentioned having worked with interpreters and translation tools before. The interpreters they worked with were not professionally trained, but rather bilingual social workers or relatives. They added that they had not used ChatGPT before, but Google Translate by writing the input, generating the audio output via the TTS function for the patients and vice versa, and in one case, they have used Gemini to enter voice input and produce text output, which the patients then read.

Next, the participants were asked about the overall communicative experience during the simulations. P1 stated that the consultation with the human interpreter felt more natural, comfortable and trustworthy, while ChatGPT did not instil trust. P2 noted that the human-mediated conversation was pleasant and fluid and did not require additional attention effort. ChatGPT, on the other hand, spoke in a fragmented manner, with noticeable pauses at the start of the turn, disrupting the natural flow and using an unfitting tone. The healthcare provider had a similar response, agreeing that the conversations with the human interpreter were smoother, more reliable and more pleasant, and ChatGPT's pauses created a feeling of insecurity whether the content was registered or would be translated, but they also noted that the consultations went well overall. The fact that the interpreter was regarded as more reliable could be linked to the ability of the interpreter to also display understanding verbally or nonverbally. In C1 and C3, there were also instances of acknowledgement displayed by the interpreter, which reassured the participants that the information was registered. Additionally, the interpreter was also taking notes, which gave the others the visual reassurance that they are being followed and listened to, and no information is getting lost.

When asked to what extent they thought they understood the patient's health complaints and concerns, the healthcare provider admitted that it worked well in all of the consultations. Upon being asked how they thought the doctor understood their problems and concerns in both scenarios, P1 chose to answer on a scale from 1 to 10, estimating a 10/10 for the simulation with the human interpreter and 6/10 for the one with ChatGPT. In C1, everything was completely understood, whereas in C2, the patient became aware of the translation mistake

from segments 18-24 regarding the body temperature, which could have led to them not trusting the system's ability to translate the whole content.

> **P1**: For the first scenario, 10 out of 10, it understood everything I wanted to communicate. But with ChatGPT, 6 out of 10. First of all, it gave wrong information when it said I had a fever of 63 degrees. Or maybe it did not translate everything I said; it only translated certain sentences from what I said.

The human interpreter was regarded as more professional and better in terms of understanding by P2 as well. ChatGPT did not hinder understanding either, but P2 admitted that they would not be able to check it if the translation was accurate. For P2, the indicator for accurate translations was the thematically appropriate responses they received from the healthcare provider; this subjective perspective being aligned with the preponderance of initiation of relevant next contributions identified in the quantitative analysis.

> **P2**: I think the translation with the interpreter was very good. It was clear that it was in a professional setting, and there were no problems. With ChatGPT, maybe because of my scenario or how I communicated, I did not notice any problems. I also did not understand what it translated, so I cannot say whether the doctor understood or not. From what they [the doctor] said to me, I gathered that they understood, meaning it conveyed the main message, and the doctor answered each of my questions or remarks. It just seemed to me that the stuttered speech or the tone it was using seemed very strange.

Consequently, the healthcare provider had the feeling that the patients understood and could follow their questions and explanations in all the scenarios, which was reciprocated by both of the patients as well.

Upon being asked whether there were any moments when they were not sure what the other person meant, the patients did not mention any difficulty. The doctor recalled the body temperature issue mentioned above, causing them to wonder if Romania was using some other unit to measure temperature rather than Celsius degrees, but apart from that, they felt that understanding was achieved.

> **A**: Yes, once with the temperature, I was not sure if they use Fahrenheit in Romania. I think that is only in America. And I tried to ask about it, but I think it was just a mistake. So ChatGPT translated it wrong somehow, but otherwise, I always felt that we understood each other.

The healthcare provider admitted that there were moments of confusion during the consultations, without further specifying them, but considered that they were all solved successfully. P2 found the part when they were discussing which of their ankle was swollen,

confusing. This was visible in segment 65 and was solved by the interpreter through a voluntary intervention to establish clarity. This shows that confusion can happen at any time, and interventions are sometimes needed in order to quickly remedy the issue. For P1, ChatGPT's delay in starting to translate, which led to the doctor ending up speaking simultaneously and interrupting the translation, caused a feeling of confusion and insecurity that the content had not been translated entirely. This concern refers to the omission in segment 62 of C2 and is proven right. ChatGPT left out the information that the treatment will be decided after the investigations, and the question regarding the pulmonologist, which was in the end repeated by the doctor.

Next, the participants evaluated which interpreting method was more useful. The healthcare provider identified advantages of both methods, such as ChatGPT's convenience of access and the ability to interact with the interpreter and ask questions. They said that both methods are preferred to not being able to communicate with a patient at all, but the human interpreter would be the preferred option.

The patients clearly sided with the human interpreter, readdressing the confidence and trustworthiness that they instil. P2 referred to specific cases in which patients may be in pain, not able to concentrate, or speak clearly and coherently. In their opinion, the interpreter would possess a better ability to assess the situation and understand the messages from context.

All participants agreed that the human interpreter conveyed their messages better, but mentioned different aspects when asked what they thought most contributed to this. For the healthcare provider, the most important aspect was the ability to interact with the interpreter and clarify ambiguous situations. This is clearly visible when looking at the evidence of understanding displayed by the doctor in the human-interpreted scenarios compared to the ChatGPT ones. The doctor displayed considerably more understanding in C1 and C3 and had direct interactions with the interpreter in segments 66-67 and 102-103 of C3, which were seen as positive voluntary interventions.

P1 and P2 appreciated the human nature of the interpreter, meaning their ability to infer meaning, despite grammatical errors, or suboptimal ways of expressing, by relying on contextual cues, interpretation, and common knowledge. P2 was especially unsure about ChatGPT's accuracy in those situations when they possibly did not phrase something correctly.

> **P2**: For example, I am not very sure when I expressed myself very... perhaps not correctly, or I used the wrong words, whether I have made myself clear through ChatGPT, and it was understood exactly what I meant, or one had to deduce it, or there

> may have been some misinterpretations. But I think the interpreter was able to do this... that is, to communicate what I wanted to say more accurately.

When asked how the interpreting method influenced the flow of conversation and the emotional atmosphere of the consultations, P1 preferred the interpreter, explaining that they did not feel ChatGPT accurately conveyed the tone of the doctor, adopting a rather aggressive intonation. This aligns with the frequency of awkward style and language register errors identified through MQM in C2. P2 felt that ChatGPT could not accurately convey the severity of their emergency or the pain level they were experiencing, and found that being able to make eye contact with the human interpreter positively impacted the natural flow of the conversation.

The doctor assessed that the conversation proceeded smoothly with the interpreter and that everything was transmitted. With ChatGPT, they reported feeling that they could not speak for too long or pause in between sentences, out of fear that it would not be recorded or translated by the system. They also noted that the interpreter does not instil this feeling, as they can simply interrupt the speaker if the turn is too long and should be translated, removing the doubt that a part of the content would get lost. Segment 66 of C1 provides an example of such a case. After segment 65, the doctor made a brief pause, which the interpreter deemed as the end of the turn and started translating. However, the doctor intended to continue speaking but refrained from doing so once the interpreter started. This behaviour illustrates effective turn-taking management, as ChatGPT showed in segment 62 of C2, which stops and omits the rendition upon interruption of other speakers. As for the emotional atmosphere of the consultation, the doctor appreciated the presence and support of the human interpreter but recognised that an additional person can make someone feel more under the pressure of judgement.

Regarding the changes in the way they would normally speak, P1 noted that they did not alter anything about their normal speech, but P2 and the healthcare provider agreed that they made a conscious effort to speak clearly, slowly, and in shorter sentences in the scenarios interpreted through ChatGPT. The doctor added that even though they did not have a strong dialect, they tried to speak standard German. These efforts can be observed in the transcripts of the simulations. In C1 and C3, there were 19 and 13 instances of using informal language from all the participants together, respectively, compared to 3 and 1, respectively, in C2 and C4. This shows the additional effort that the participants (especially A and P2) made to speak as clearly and correctly as possible.

On the topic of empathy and the ability to connect with the patient, the healthcare provider stated that neither method of interpreting impaired their effort, as the patients could hear the tone of their voice either way. Their impression was that ChatGPT's voice in German

was very unempathetic and that the interpreter could convey a similar tone to their own. The transmission of empathy was deemed important by the doctor, and their efforts to establish rapport with the patients are also visible through the repetition of the question „Ja?" [Yes?] at the end of sentences in longer explanatory segments, such as 67 in C1, 58 in C2, 93 in C3 or 37 in C4. The human-interpreted consultations allowed for longer turns for one speaker, each of the aforementioned containing 5 such examples, compared to 2 and 1 in C2 and C4, respectively.

The understanding of an interlocutor is also assessed nonverbally by a speaker through their reaction upon receiving the information. On the question of whether they noticed any surprising reaction, which could indicate a misunderstanding, P1 recalled the situation about the fever, when they understood from the healthcare provider's reaction that the message was wrongly transmitted, but P2 and the healthcare specialist did not encounter any such situations.

When asked if they became aware of any omissions or changes in the translations, P1 noted that even though there were some omissions, they would not have influenced the final diagnosis. The doctor stated that they did not notice such occurrences, since the responses they received were accurate and fitting, but they were more certain that the interpreter delivered a complete translation, as they could see them taking notes of everything that was said. P2 had the impression that there were slightly more mistakes made in C4 than in C3, which is, however, untrue, when compared to the Quality analysis of the two frameworks. As for unaltered segments, C3 had 76.67, and C4 had 88.46, and the MQM scores were identical. This opinion could be rooted in bias or a certain scepticism towards the S2S translation system, but could also be caused by the fact that in C4, ChatGPT showed considerable response latency, lowering the overall impression of quality and reliability.

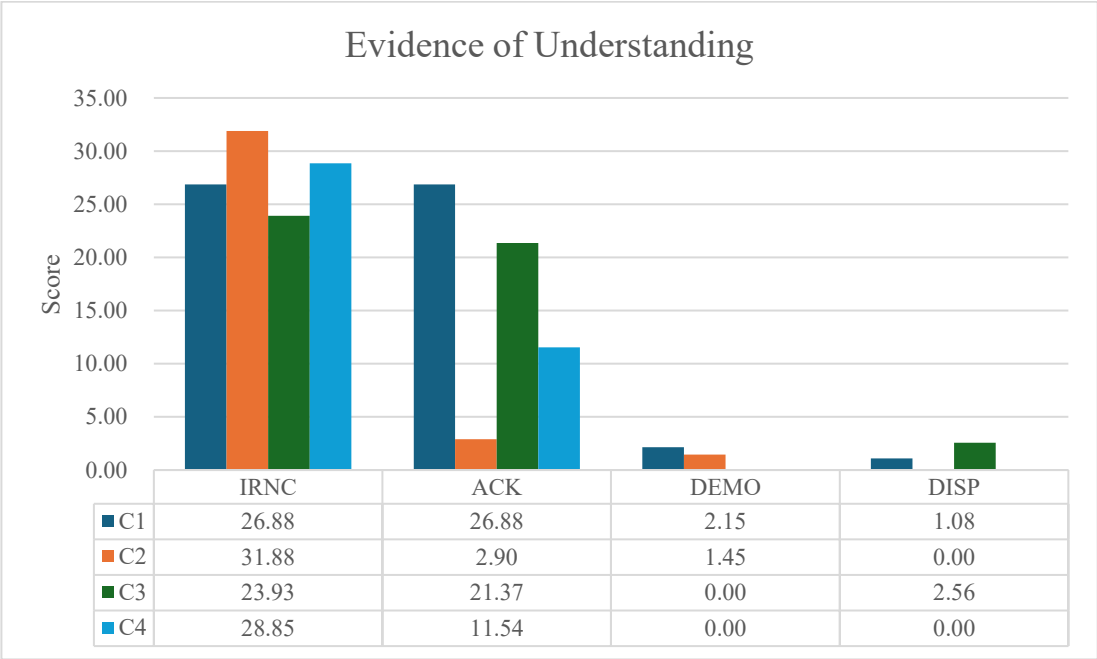
In terms of the use of medical terminology, the participants did not notice any difference, but the doctor noted that the terms were more obvious and recognisable in the ChatGPT scenarios. They recognised the terminology in Romanian as well, by associating it with the Latin terms. A recognisable trend is that the interpreter tended to rephrase or simplify the medical terminology to facilitate understanding, such as in segments 20 of C1 and 104 and 52 of C3, whereas ChatGPT used the precise medical terms, such as in segments 15 of C2 and 24 of C4. There was, however, an instance where the medical term „boli cronice" [chronic diseases] did not convey the exact meaning in the target language as the source utterance, but it went unnoticed by the participants and did not cause any harm or understanding difficulties.

Upon noticing confusion, the healthcare provider and P1 noted that they successfully repaired the problems by repeating the questions that did not get translated, and then they

received the expected answer. P2 added that it was not necessarily errors, but the awkward atmosphere created by the fragmented and delayed speech of ChatGPT that impaired the flow of the conversation.

The last topic discussed in the interview with the healthcare provider was their addressee ambiguity in the human-mediated consultations. The doctor admitted to being uncertain whether to address the patient directly or speak in the third person through the interpreter to them. In contrast, communication via ChatGPT urges them to address the patient directly, simplifying the interaction and reducing ambiguity. This uncertainty was demonstrated in segments 15 and 31 of C1, when the doctor asked the interpreter about the patient in the third person. Interestingly, ChatGPT once made the reverse action in segment 8 of C4, by talking about the patient in the third person to the doctor.

8.4. Discussion Understanding



	IRNC	ACK	DEMO	DISP
C1	26.88	26.88	2.15	1.08
C2	31.88	2.90	1.45	0.00
C3	23.93	21.37	0.00	2.56
C4	28.85	11.54	0.00	0.00

Figure 6. Evidence of Understanding across all consultations

When looking at all evidence of understanding scores in Figure 6, it becomes clear that the dominant type of evidence in all the consultations was the initiation of the relevant next contribution. In the simulations interpreted by ChatGPT, however, these scores were more dominant compared to the ones where the interpreter was human. This may be due to the increased representation of acknowledgement in C1 and C3. The contribution of the human interpreter to show acknowledgement also played a role in the higher acknowledgement scores of the human-interpreted simulations. Demonstrations were only present in C1 and C2, but only scored 2.15 and 1.45, respectively, while displays were missing completely from the S2S system scenarios.

In several cases, participants exhibited understanding before any interpreting occurred, likely because certain lexical items are phonologically similar between Romanian and German or are shared borrowings from English. In C2 (segments 7, 57, 60, and 69), and C3 (segments 101, 103, 116, and 117), understanding was displayed silently, or through the initiation of the next relevant contribution, bypassing the interpretation altogether.

Following the typical doctor-patient interactional pattern described by Hofer-Falk (2023), the physician led the consultations, asking questions, while the patient answered with relevant contributions. As the healthcare provider continued the examination in a natural way by making further inquiries about other topics without initiating repairs, the assumption is that

the patient's contributions were silently accepted and understood by the doctor. As Hofer-Falk (2023) recognises, this orderly alteration of speakers establishes internal interactional coherence, which facilitates the understanding while not guaranteeing it.

The quantitative analysis of the perceived understanding showed that the participants exhibited more frequent and stronger evidence of understanding in consultations mediated by the human interpreter than in those with ChatGPT. This pattern was also visible in the interview data. The qualitative analysis revealed that the participants had a general feeling of understanding throughout all the consultations. Building on the view that rapport and empathy are key aspect of doctor-patient understanding associated with better patient outcomes (Derksen et al. 2013; Iglesias Fernandez 2010), participants consistently reported that the human interpreter sounded more empathetic, whereas ChatGPT did not. P2 noted that ChatGPT failed to convey urgency and severity adequately, which diminished the perceived alignment with the physician's intent. Across all participants, empathy was regarded as important for understanding, and both patients and the doctor considered that the physician's empathetic tone was audible and inferable through prosody. Additionally, P2 positively appreciated the contribution of eye contact with the human interpreter to the conversation flow.

The duration of the consultations was similar (C1 – 11:46, C2 – 8:45, C3 – 9:18, C4 – 9:40), but when it comes to the segment count, C1 and C3 had 93 and 117 segments, respectively, while C2 and C4 had 69 and 52, respectively. Both the clinician and one patient noted that, in ChatGPT-mediated consultations, they intentionally produced shorter sentences because they doubted the system would record the full content. Considering that consultation durations were similar, yet segment counts were substantially lower with ChatGPT, the interactional data from the interview align with these reports and suggest increased response latency for ChatGPT. This eroded confidence in the tool, weakening rapport and causing turn-taking difficulties. Specifically, long pauses resulted in speakers repeating questions they assumed were not recorded; this often coincided with the system's delayed start, which caused start–stop overlaps.

This uncertainty and lack of trust in the tool's ability to capture the entire content of a turn contrasts with the confidence evoked by the interpreter, who voiced acknowledgement that the content had been understood, and, as the healthcare provider observed, the notes of the interpreter were visible, so the whole process of registering the content was visible for the participants.

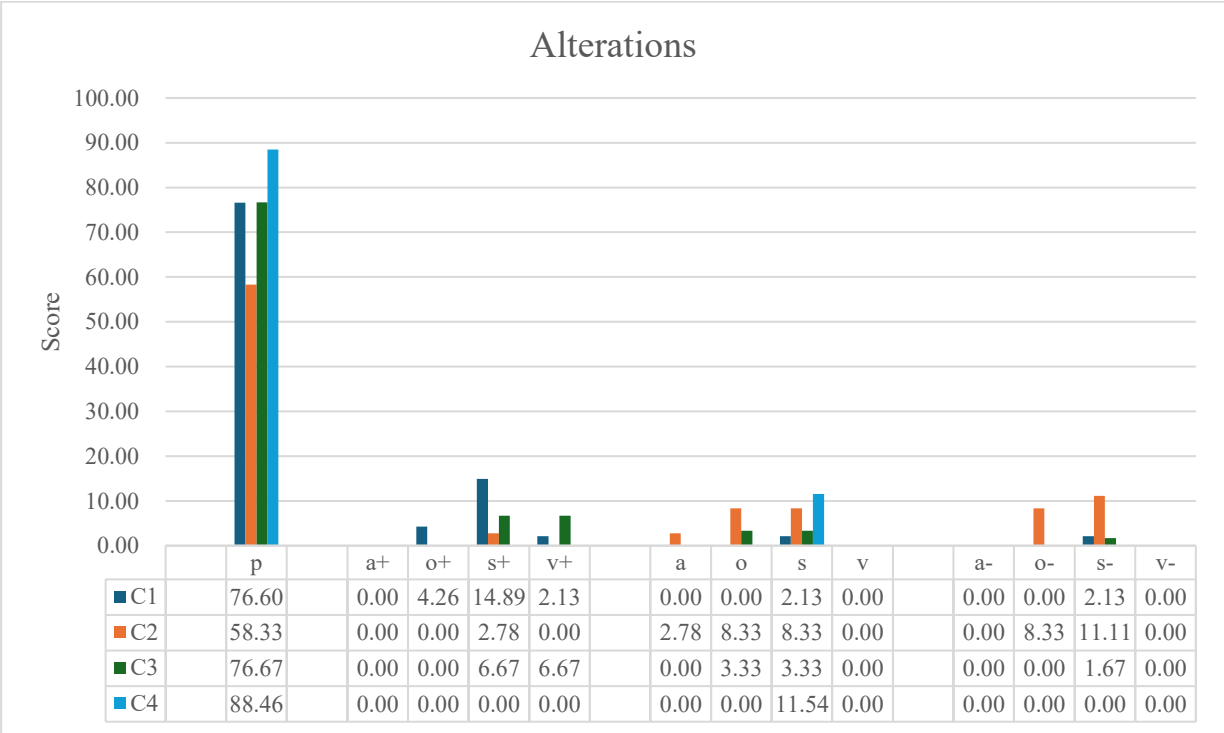
Informal speech was more prevalent in the human-mediated consultations and even occurred for the interpreter. This was, however, not considered an error, as it was interactionally

fitting to the scenario. On the other hand, ChatGPT was formal, and the physician noted that they deliberately tried to speak in standard German to avoid confusing the system. In turn, P2 expressed their concern regarding the system's ability to comprehend their imperfect expressions, which strengthened their preference for the human interpreter

Another factor which decreased rapport in the ChatGPT scenarios was the unnatural prosody and unfitting tone evidenced through the awkward style mistakes and the comments of the patients in the interview. P1 criticised the aggressive tone of the system, which is aligned with the less polite rendering in segment 22 of C2. In addition, especially in P1's scenario, the TTS pronunciation sometimes adopted a German-like rolled 'r', perceived as odd and amusing by the participants.

Altogether, understanding was reached in all the scenarios, and even the instances which posed an obstacle to immediate understanding were successfully solved. Despite this, participants voiced a preference for the human interpreter when it comes to facilitating understanding, especially through trust, empathy and rapport.

8.5. Discussion Quality



	p	a+	o+	s+	v+	a	o	s	v	a-	o-	s-	v-
C1	76.60	0.00	4.26	14.89	2.13	0.00	0.00	2.13	0.00	0.00	0.00	2.13	0.00
C2	58.33	0.00	0.00	2.78	0.00	2.78	8.33	8.33	0.00	0.00	8.33	11.11	0.00
C3	76.67	0.00	0.00	6.67	6.67	0.00	3.33	3.33	0.00	0.00	0.00	1.67	0.00
C4	88.46	0.00	0.00	0.00	0.00	0.00	0.00	11.54	0.00	0.00	0.00	0.00	0.00

Figure 7. Alterations across all consultations

With respect to quality, Figure 7 depicts the analysis under Hamai et al. (2024), and Figure 8 shows the final MQM scores. Looking at the scores of all consultations in Figure 7, the first observation is that the total scores of unaltered, i.e. accurate segments are 76.60 and 76.67 for the human-interpreted simulations and 58.33 and 88.46 for the ones mediated by ChatGPT. The scores for C1 and C3 are very similar, while the ones for C2 and C4 differ substantially. Overall, the human interpreter managed to produce consistently more accurate segments. A greater difference between human and machine is visible when considering the total amounts of positive alterations in addition to the accurate translations. Here, it becomes clearer that the human interpreter significantly improved the quality of the consultations, scoring 21.28 for C1 and 13.34 for C3, when it comes to the positive alterations, against 2.78 for the S2S translation system.

As for the clinically significant and insignificant negative alterations: omissions and substitutions occurred most frequently for both translation methods. There was only one case of an insignificant addition in C2 and no cases of significant or insignificant negative voluntary interventions.

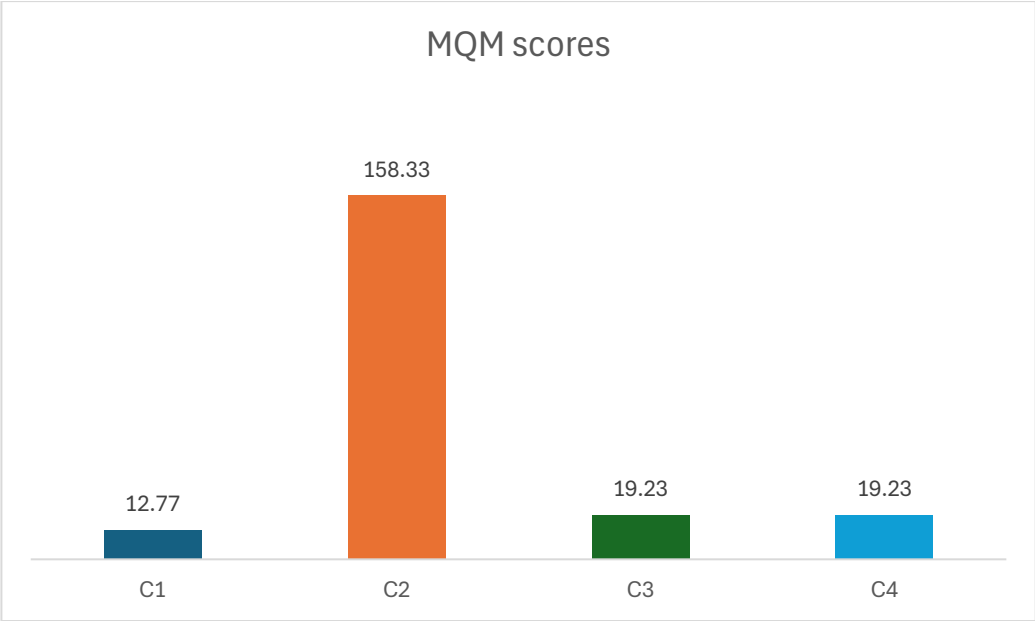


Figure 8. MQM score of each consultation

Figure 8 indicates similar quality for C1, C3 and C4 (12.77, 19.23, and 19.23), whereas C2 scores notably poorer. The latter exhibits by far the most negative alterations and the gravest errors, mainly because of the presence of numbers in the source utterances, which considerably lowered the performance of ChatGPT. Although P2 participated in C3 and C4, where both the interpreter and ChatGPT received similarly high scores under Hamai et al. (2024) and the same MQM score, they judged ChatGPT substantially poorer. This discrepancy may reflect a trust gap caused by the inability to obtain confirmation whether the system captured and translated their speech accurately. This fostered scepticism, which likely led to a biased perceived quality.

As this thesis focuses on interpretation instead of translation, renditions that differed in form from the source utterance, but brought across the entire message with the intended purpose, were not penalised. Although the two chosen quality assessment frameworks serve different purposes and focus on slightly different aspects of the translation or interpreting process, the negative alterations from Hamai et al. (2024) could be compared to the error types identified through the MQM. Negative, insignificant alterations were marked as minor issues, while significant ones represented either major or critical MQM errors. Additions and omissions in both frameworks refer to instances where content is added or removed from the source utterance. Substitutions could be compared to mistranslations, as well as language register and grammar issues in MQM. There was, however, one category of MQM errors related to style, which had no correspondent in Hamai et al. (2024). The interpreter made only one style error in segment 2, whereas ChatGPT produced numerous style errors in C2 and C4. Those

types of errors were not severe enough to impede the understanding or meaning of the source speech in any way and were consequently deemed as minor under MQM, while not even being considered errors in Hamai et al. (2024), but contributed to the overall impression of the patients that the tone and prosody of the system were not right.

Overall, the fact that the human interpreter delivered consistently high quality across consultations, whereas ChatGPT showed variability. This inconsistency undermines the system's reliability, even when it outperformed the human interpreter in C4.

8.6.Additional Findings

The post-hoc transcript audit revealed mismatches between the ASR, MT and TTS components in the pipeline of ChatGPT. Amongst the identified cases were incorrect or missing transcripts paired with acceptable audio, and vice versa, or even complete hallucinations in the transcripts, such as in segment 53 "Do not forget to like, leave a comment, and share this video on other social media platforms":

53.	P1.2	Transcript Nu uitați să dați like, să lăsați un comentariu și să distribuiți acest material video pe alte rețele sociale. Voice Mmm, da, adică mă doare în partea dreaptă a pieptului și mă doare și mai tare când respir adânc. Se agravează durerea când respir adânc.
54.	ChatGPT	Ja, es schmerzt auf der rechten Seite der Brust, und der Schmerz wird noch stärker, wenn ich tief einatme, der Schmerz verschlimmert sich also beim tiefen Atmen.

Excerpt from Appendix 26. Example of hallucination

Because these inconsistencies emerged only during the transcript analysis and were not part of the participants' experiences, they will be treated as additional observations.

Another notable observation is drawn from the physician's interview response, which highlights that while human interpreters provide valuable support, their presence as an additional participant in the interaction can sometimes make the other parties feel judged. In comparison, ChatGPT may offer similar linguistic assistance without introducing the social pressure associated with a human intermediary; however, it lacks the capacity to offer the emotional support that human interpreters provide.

9. Conclusion

The goal of this thesis was to determine the role of language technology – specifically, an LLM-based system (ChatGPT) capable of translating and interpreting, in contemporary multilingual doctor-patient communication. More precisely, this thesis examined the perception of understanding and the quality through the following research questions:

> RQ1: How does S2S translation via ChatGPT compare to human interpreting when assessing the perceived understanding of healthcare providers and patients during medical consultations in German and Romanian?
>
> RQ2: What differences in quality emerge between human interpreters and S2S translation via ChatGPT when analysed through the framework proposed by Hamai et al. (2024) and MQM?

To investigate these aspects, four simulated medical emergency room consultations with two Romanian-speaking patients and one German-speaking healthcare provider were organised, and each patient got to converse with the healthcare provider once with the help of a professional interpreter and once via the S2S translation function of ChatGPT. To gather their impressions of the consultations, a group interview with the patients and a separate one with the doctor were held after the simulations. The focus of this qualitative and quantitative research was on the perception of understanding of the patients and the healthcare provider, as well as on the quality of the interpretations produced by ChatGPT against those of a professional interpreter.

The thesis discusses the conduit and triadic models to parallel the transmission of information in interactions involving ChatGPT and a human interpreter, the latter of which relies on interactional practices such as turn management to enhance the communication (Downie 2020; Pöchhacker 2022; Roy 1992). Clark & Schaefer's (1989) grounding framework was examined to demonstrate the importance of exhibiting understanding in discourse. This contrast, consisting in the ability of the interpreter to also display understanding, suggests why the participants regarded the human interpreter as more trustworthy and their performance as more qualitative, even when formal quality metrics were similar. Additionally, this thesis situates these findings within the context of healthcare interpreting, where effective doctor-patient communication, strengthened through empathy and rapport, leads to positive health outcomes (Hofer-Falk 2023; Iglesias Fernandez 2010; Stewart 1995). The interventions of the

human interpreter, clearly visible through the analysis of the positive alterations, increase the overall quality of the interpretation and align with the triadic model, which assumes that interpreters are active participants in the conversation and contribute to reaching the goal of the involved parties in need of the interpretation to communicate. Gile's (2009) Effort Model for consecutive interpreting served to better understand where error potential lies for human interpreters and to understand the cognitive processing required compared to the S2S translation system pipeline of ASR – MT – TTS (Tan 2023).

Starting with RQ1, all participants reported mutual understanding in all consultations, independent of the interpretation method. The quantitative analysis of the evidence of understanding based on Clark & Schaefer (1989) confirms this, as despite the human-mediated scenarios containing stronger and more frequent evidence, nearly every turn across all simulations consisted either of mostly relevant next contributions to the previous turn, or other types of evidence of understanding. The consultations followed the typical doctor-patient communication pattern described by Hofer-Falk (2023), reinforcing the intrinsic coherence of the discussion and favouring understanding. Although mutual understanding was reached, all the participants voiced a preference for the human interpreter, citing different reasons such as greater trust, and superior handling of context and pragmatics, and the accuracy of conveying empathy, urgency and tone.

Moving on to RQ2, the perceived quality of the participants in favour of the human interpreter was not supported by the quantitative data in each consultation. The analysis of alterations based on Hamai et al. (2024) showed a total score of unaltered, i.e. accurate segments of 76.60 (C1) and 76.67 (C3) for the human-interpreted simulations and 58.33 (C2) and 88.46 (C4) for the ones mediated by ChatGPT, indicating a comparable quality for C1 and C3, higher quality for C4, and lower quality for C2. As this framework accounts for the positive influence that interactional decisions of the interpreter can have on an interpretative act, the significant added value of the human interpreter becomes clear when taking the positive alterations into account as well.

Viewed through MQM, the quality of C1, C3 and C4 ranged from good to very good, while C2 achieved a significantly poorer score. The participants regarded the human interpretation as higher in quality, but this is contradicted by the comparison between C3 and C4, where ChatGPT surpassed the human interpreter. This divergence can be consequent to a lack of trust in the capabilities of the system, which creates a biased perception of quality.

ChatGPT yielded the lowest scores across both frameworks in C2. This was likely due to the higher density of numbers and the system's difficulties in processing them, which led to

severe mistranslations and omissions. This decrease in performance undermines the reliability of the system compared to the human interpreter, who maintained a consistent quality throughout the consultations. This variability has serious implications in domains where human lives are at stake, such as healthcare, where dependable accuracy is crucial.

As there is no established quality evaluation framework for community interpreting (Hofer-Falk 2023), this thesis offers insight into the aspects to be considered when assessing interpreters in such contexts. By using both the MQM and the Hamai et al. (2024) frameworks, this thesis presents a more holistic evaluation of interpreting quality, where linguistic and accuracy, but also interactional effectiveness, are important. This is especially relevant when considering whether or in which scenarios S2S translation systems could replace human interpreters entirely in the future.

With regard to the initial assumptions, the findings of this study provide support for both. The first assumption, namely that perceived understanding would be higher in interactions involving a human interpreter compared to the S2S translation system, is largely confirmed. While S2S technology proved to be a fast and functional tool for information exchange and even outperformed the accuracy of the interpreter in one consultation, participants generally reported a higher level of understanding, trust, comfort and communicative clarity when interacting with a human interpreter.

The second assumption, that participants would consciously modify their speech when interacting with ChatGPT, is also supported by the data. Participants appeared to adapt their speech by shortening their utterances and adopting a standard German or Romanian, suggesting an increased awareness of the need to facilitate machine processing, but also a lack of trust in the ability of the system. This is corroborated by the comparison between standard and informal speech across the consultations, as well as by the participants' responses during the interviews.

Finally, given the context of healthcare interpreting and the use of personal information, it is important to reflect on the ethical implications of using tools such as ChatGPT. While there are regulations in place, such as the GDPR or the EU AI Act, companies developing AI systems may collect user data to train their models, which raises potential security risks, meaning that the responsibility to provide data safety lies mostly with the users. Professional interpreters are bound by codes of ethics to respect the confidentiality of their clients, and, through their personal and professional judgement as active participants, they can make decisions to safeguard patients' well-being and promote high-quality care. Consequently, the use of AI-based S2S translation systems may have broader implications for data protection and trust in healthcare settings.

Limitations and Future Research

The present study is subject to several limitations. For confidentiality reasons, the investigation in a natural environment (a real doctor-patient consultation) was not possible, so simulations were done instead. Since no real health concerns were at stake, the interaction lacked the urgency and emotional involvement an authentic situation would have. This may have impacted the way the participants communicated and the interpreter acted, which limits the validity of the findings.

A further constraint is the small number of participants, which provided a limited dataset. This hinders the potential of generalising the findings of the thesis, since the study is not representative.

As the analysis was done at the segment level, a methodological challenge appeared due to the nature of consecutive interpreting, which does not always allow for clear-cut segmentation, since overlapping or fragmented speech can occur. Moreover, the choice to conduct the analysis at the segment level poses the potential risk of error underrepresentation in segments which contained multiple instances of the same error.

Since interpreting is a complex, dynamic process, future research could explore the integration of linguistic and interactional dimensions to assess quality in a comprehensive manner. Another potential avenue for investigation is the shift of translation and interpreting systems towards the triadic model. Even though MT tools were regarded as “black boxes” (Horváth 2021: 178) in the past, recent technological advancements have opened the door for them to be integrated as participants in a triadic constellation through prompts and user interactions.

Bibliography

Allison, Abigail & Hardin, Karol (2020). Missed Opportunities to Build Rapport: A Pragmalinguistic Analysis of Interpreted Medical Conversations with Spanish-Speaking Patients. *Health Communication*,*35* (4), 494–501. DOI: 10.1080/10410236.2019.1567446.

ALPAC (1966). *Language and Machines: Computers in Translation and Linguistics*. Washington, D.C.: National Academy of Sciences, National Research Council. 9547.

BDÜ (2023). FAQ Gesetzesvorhaben Dolmetschen im Gesundheitswesen. BDÜ. https://bdue.de/fileadmin/files/PDF/Positionspapiere/BDUe_FAQ_Gesetzesvorhaben_Dolmetschen_im_Gesundheitswesen_2023.pdf.

Bischoff, Alexander & Loutan, Louis (2004). Interpreting in Swiss hospitals. *Interpreting. International Journal of Research and Practice in Interpreting*,*6* (2), 181–204. DOI: 10.1075/intp.6.2.04bis.

Bot, Hanneke (2005). *Dialogue Interpreting in Mental Health*. Leiden Boston: BRILL. DOI: 10.1163/9789004458574.

Brandenberger, Julia; Stedman, Ian; Stancati, Noah; Sappleton, Karen; Kanathasan, Sarathy; Fayyaz, Jabeen & Singh, Devin (2025). Using artificial intelligence based language interpretation in non-urgent paediatric emergency consultations: a clinical performance test and legal evaluation. *BMC Health Services Research*,*25* (1), 138. DOI: 10.1186/s12913-025-12263-1.

Chen, Kang; Zhou, Xiuze; Lin, Yuanguo; Feng, Shibo; Shen, Li & Wu, Pengcheng (2025). A survey on privacy risks and protection in large language models. *Journal of King Saud University Computer and Information Sciences*,*37* (7), 163. DOI: 10.1007/s44443-025-00177-1.

Clark, Herbert H. & Schaefer, Edward F. (1989). Contributing to Discourse. *Cognitive Science*,*13* (2), 259–294. DOI: 10.1207/s15516709cog1302_7.

DeepL SE (2025). *DeepL Translator*. https://www.deepl.com (Accessed: 19.1.2026).

Deng, Mingjun & Fan, Xianming (2025). Post-Editing Efficiency And Quality Assessment: A Comparative Analysis Of Google Translate And DeepL. *IOSR Journal of Humanities and Social Science*,*30* (1), 10–24. DOI: 10.9790/0837-3001021024.

Derksen, Frans; Bensing, Jozien & Lagro-Janssen, Antoine (2013). Effectiveness of empathy in general practice: a systematic review. *British Journal of General Practice*,*63* (606), e76–e84. DOI: 10.3399/bjgp13X660814.

Dorner, Britta; Kuznik, Anna; Orego Carmona, David & Zwischenberger, Cornelia (2025). Surveys an interviews. In: Rojo López, Ana María, & Muñoz Martín, Ricardo (Eds) *Research Methods in Cognitive Translation and Interpreting Studies*. Amsterdam: John Benjamins Publishing Company. Vol. 10. 69–91. DOI: 10.1075/rmal.10.

Downie, Jonathan (2020). *Interpreters vs machines: can interpreters survive in an AI-dominated world?* London ; New York: Routledge, Taylor & Francis Group.

EU Regulation (EU) 2016/679 (General Data Protection Regulation). (2016). https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32016R0679.

EU Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on harmonised rules on artificial intelligence (Artificial Intelligence Act). (2024). https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32024R1689.

Euromonitor International (2025). *Analgesics in Romania*. Euromonitor International.

Fantinuoli, Claudio (2018). Interpreting and Technology: The Upcoming Technological Turn. In: *Interpreting and technology*. Zenodo. 1–12. DOI: 10.5281/ZENODO.1493281.

Gile, Daniel (2009). *Basic Concepts and Models for Interpreter and Translator Training: Revised edition*. Amsterdam: John Benjamins Publishing Company. http://ebookcentral.proquest.com/lib/univie/detail.action?docID=622263.

Google (2025).a *Google Gemini*. https://gemini.google.com (Accessed: 19.1.2026).

Google (2025).b *Google Translate*. https://translate.google.com (Accessed: 19.1.2026).

Guo, Wei; Guo, Xun; Huang, Junkang & Tian, Sha (2024). Modeling listeners' perceptions of quality in consecutive interpreting: a case study of a technology interpreting event. *Humanities and Social Sciences Communications*,*11*, 985. DOI: 10.1057/s41599-024-03511-6.

Hale, Sandra (2014). Interpreting culture. Dealing with cross-cultural issues in court interpreting. *Perspectives*,*22* (3), 321–331. DOI: 10.1080/0907676X.2013.827226.

Hamai, Taeko; Nagata, Ayako; Ono, Naoko; Nishikawa, Hiroaki & Higashino, Sadanori (2024). Evaluating a conceptual framework for quality assessment of medical interpretation. *Patient Education and Counseling*,*123*. DOI: 10.1016/j.pec.2024.108233.

Hatim, Basil & Mason, Ian (1997). *The translator as communicator*. London: Routledge. Vol. 1.

Herrmann-Werner, Anne; Loda, Teresa; Zipfel, Stephan; Holderried, Martin; Holderried, Friederike & Erschens, Rebecca (2021). Evaluation of a Language Translation App in an Undergraduate Medical Communication Course: Proof-of-Concept and Usability Study. *JMIR mHealth and uHealth*,*9* (12), e31559. DOI: 10.2196/31559.

Hidayati, Niswatin Nurul & Nihayah, Dewi Hidayatun (2024). Google Translate, ChatGPT or Google Bard AI: A Study toward Non-English Department College Students' Preference and Translation Comparison. *Inspiring: English Education Journal*,*7* (1), 14–33. DOI: 10.35905/inspiring.v7i1.8821.

Hofer-Falk, Gertrud (2023). *Gedolmetschte Ärzt:innen-Patient:innen-Gespräche: Phänomene und Probleme aus gesprächsanalytischer und aus dolmetschwissenschaftlicher Perspektive*. Gunter Narr Verlag. 1st edn. DOI: 10.24053/9783823395546.

Horváth, Ildikó (2021). Speech translation vs. Interpreting. *Language Studies and Modern Humanities*,*3* (2), 174–187. DOI: 10.33910/2686-830X-2021-3-2-174-187.

Hua, Shangying; Jin, Shuangci & Jiang, Shengyi (2024). The Limitations and Ethical Considerations of ChatGPT. *Data Intelligence*,*6* (1), 201–239. DOI: 10.1162/dint_a_00243.

Hudelson, Patricia & Chappuis, François (2024). Using Voice-to-Voice Machine Translation to Overcome Language Barriers in Clinical Communication: An Exploratory Study. *Journal of General Internal Medicine*,*39* (7), 1095–1102. DOI: 10.1007/s11606-024-08641-w.

Hutchins, John W. (2023). Machine Translation: History of Research and Applications. In: Sin-wai, Chan (Ed.) *Routledge Encyclopedia of Translation Technology*. London: Routledge. 2nd edn. 128–144. DOI: 10.4324/9781003168348.

Iglesias Fernandez, Emilia (2010). Verbal and nonverbal concomitants of rapport in health care encounters: implications for interpreters. *The Journal of Specialised Translation*,(14), 216–228. DOI: 10.26034/cm.jostrans.2010.586.

IMIA (2006). *IMIA Code of Ethics for Medical Interpreters*. https://www.imiaweb.org/code/ger.asp?utm_source=chatgpt.com (Accessed: 8.2.2026).

Ishida, Toru; Murakami, Yohei & Bramantoro, Arif (2025). *Impact of Large Language Models on Conversational Translation: Towards Translation Agents*. In: *2025 7th International Conference on Applied Computational Intelligence in Information Systems (ACIIS)*.

Bandar Seri Begawan, Brunei Darussalam: IEEE. 1–6. DOI: 10.1109/ACIIS66255.2025.11402976.

iTranslate GmbH (2025). *iTranslate Converse*. https://itranslate.com/converse (Accessed: 21.8.2025).

Ji, Meng; Bouillon, Pierrette & Seligman, Mark (2023). *Translation Technology in Accessible Health Communication*. Cambridge University Press. 1st edn. DOI: 10.1017/9781108938976.

Jia, Ye; Ramanovich, Michelle Tadmor; Remez, Tal & Pomerantz, Roi (2022). Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*,*162*, 10120–10134.

Jia, Ye; Weiss, Ron J.; Biadsy, Fadi; Macherey, Wolfgang; Johnson, Melvin; Chen, Zhifeng & Wu, Yonghui (2019).a Direct Speech-to-Speech Translation with a Sequence-to-Sequence Model. *Interspeech 2019*,1123–1127. DOI: 10.21437/Interspeech.2019-1951.

Jia, Ye; Weiss, Ron J.; Biadsy, Fadi; Macherey, Wolfgang; Johnson, Melvin; Chen, Zhifeng & Wu, Yonghui (2019).b *Direct speech-to-speech translation with a sequence-to-sequence model*.Paper presented at the 'Proceedings of Interspeech 2019'. 1123–1127. DOI: 10.48550/arXiv.1904.06037.

Johnson, Melvin; Schuster, Mike; Le, Quoc V.; Krikun, Maxim; Wu, Yonghui; Chen, Zhifeng; Thorat, Nikhil; Viégas, Fernanda; Wattenberg, Martin; Corrado, Greg; Hughes, Macduff & Dean, Jeffrey (2017). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. arXiv. DOI: 10.48550/arXiv.1611.04558.

Khoong, Elaine C.; Steinbrook, Eric; Brown, Cortlyn & Fernandez, Alicia (2019). Assessing the Use of Google Translate for Spanish and Chinese Translations of Emergency Department Discharge Instructions. *JAMA Internal Medicine*,*179* (4), 580. DOI: 10.1001/jamainternmed.2018.7653.

Kitano, Hiroaki (1994). *Speech-to-Speech Translation: A Massively Parallel Memory-Based Approach*. Boston, MA: Springer US. DOI: 10.1007/978-1-4615-2732-9.

Leanza, Yvan (2007). Roles of community interpreters in pediatrics as seen by interpreters, physicians and researchers. In: Pöchhacker, Franz, & Shlesinger, Miriam (Eds) *Healthcare Interpreting: Discourse and Interaction*. Amsterdam; Philadelphia: John Benjamins. 11–34.

Leite, Fernando Ochoa; Cochat, Catarina; Salgado, Henrique; Da Costa, Mariana Pinto; Queirós, Marta; Campos, Olga & Carvalho, Paulo (2016). Using Google Translate^© in the hospital: A case report. *Technology and Health Care*,*24* (6), 965–968. DOI: 10.3233/THC-161241.

Li, Xinchen (2024). Comparison of Translation Quality between Large Language Models and Neural Machine Translation Systems: A Case Study of Chinese-English Language Pair. *International Journal of Education and Humanities*,*4* (2), 121–128. DOI: 10.58557/(ijeh).v4i2.213.

Lommel, Arle (2014). *Multidimensional Quality Metrics (MQM) Specification*. https://tranquality.info/mqm/mqm-spec-2014-02-14.html (Accessed: 2.2.2026).

Lommel, Arle; Gladkoff, Serge; Melby, Alan; Wright, Sue Ellen; Strandvik, Ingemar; Gasova, Katerina; Vaasa, Angelika; Benzo, Andy; Sparano, Romina Marazzato; Foresi, Monica; Innis, Johani; Han, Lifeng & Nenadic, Goran (2024). The Multi-Range Theory of Translation Quality Measurement: MQM scoring models and Statistical Quality Control. arXiv. DOI: 10.48550/arXiv.2405.16969.

Lommel, Arle; Uszkoreit, Hans & Burchardt, Aljoscha (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality

Metrics. *Tradumàtica Tecnologies de La Traducció*,(12), 455–463. DOI: 10.5565/rev/tradumatica.77.

Lyu, Chenyang; Du, Zefeng; Xu, Jitao; Duan, Yitao; Wu, Minghao; Lynn, Teresa; Aji, Alham Fikri; Wong, Derek F. & Wang, Longyue (2024). A Paradigm Shift: The Future of Machine Translation Lies with Large Language Models. *Machine Translation*,1339–1352. DOI: https://doi.org/10.63317/3dqcen72ckgz.

Merlini, Raffaela & Favaron, Roberta (2007). Examining the "voice of interpreting" in speech pathology. In: Pöchhacker, Franz, & Shlesinger, Miriam (Eds) *Healthcare Interpreting: Discourse and Interaction*. Amsterdam; Philadelphia: John Benjamins. 101–137.

Microsoft Corporation (2025). *Microsoft Translator*. https://www.microsoft.com/en-us/translator/ (Accessed: 21.8.2025).

Moniz, Helena & Parra Escartín, Carla (ed.) (2023). *Towards Responsible Machine Translation: Ethical and Legal Considerations in Machine Translation*. Cham: Springer International Publishing. Vol. 4. DOI: 10.1007/978-3-031-14689-3.

Moser-Mercer, Barbara (1997). Process models in simultaneous interpretation. In: Hauenschild, Christa, & Heizmann, Susanne (Eds) *Machine Translation and Translation Theory*. Berlin, New York: De Gruyter Mouton. DOI: 10.1515/9783110802474.3.

NCIHC (2004). National Code of Ethics for Interpreters in Health Care. https://www.ncihc.org/assets/Accessible-files/NCIHC%20National%20Code%20of%20Ethics%20accessible.pdf.

Norfolk, Tim; Birdi, Kamal & Walsh, Deirdre (2007). The role of empathy in establishing rapport in the consultation: a new model. *Medical Education*,*41* (7), 690–697. DOI: 10.1111/j.1365-2923.2007.02789.x.

Obeidat, Mohammed M.; Haider, Ahmad S.; Tair, Sausan Abu & Sahari, Yousef (2024). Analyzing the Performance of Gemini, ChatGPT, and Google TranslateinRendering English Idioms into Arabic. *FWU Journal of Social Sciences*,*18* (4), 1–18. DOI: 10.51709/19951272/Winter2024/1.

OECD (2019). *Talent Abroad: A Review of Romanian Emigrants*. OECD. DOI: 10.1787/bac53150-en.

Olsavszky, Victor; Bazari, Mutaz; Dai, Taieb Ben; Olsavszky, Ana; Finkelmeier, Fabian; Friedrich-Rust, Mireen; Zeuzem, Stefan; Herrmann, Eva; Leipe, Jan; Michael, Florian Alexander; Westernhagen, Hans Von & Ballo, Olivier (2025). Digital Translation Platform (Translatly) to Overcome Communication Barriers in Clinical Care: Pilot Study. *JMIR Formative Research*,*9*, e63095. DOI: 10.2196/63095.

Ong, L. M. L.; De Haes, J. C. J. M.; Hoos, A. M. & Lammes, F. B. (1995). Doctor-patient communication: A review of the literature. *Social Science & Medicine*,*40* (7), 903–918. DOI: 10.1016/0277-9536(94)00155-M.

Open AI (2026). *Entdecke GPT-5.4*. https://openai.com/de-DE/index/introducing-gpt-5-4/ (Accessed: 31.3.2026).

OpenAI (2022). *Introducing ChatGPT*. https://openai.com/index/chatgpt/ (Accessed: 6.1.2026).

OpenAI (2025).a *ChatGPT*. [Generative AI chat]. https://chat.openai.com/.

OpenAI (2025).b *Introducing GPT-5.2*. https://openai.com/index/introducing-gpt-5-2/ (Accessed: 10.1.2026).

OpenAI (2026). *How your data is used to improve model performance*. https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance?utm_source=chatgpt.com (Accessed: 10.2.2026).

ÖVGD (2025). *ÖVGD – Über den ÖVGD*. https://oevgd.at/de/uber-den-oevgd (Accessed: 8.2.2026).

Pöchhacker, Franz (1999). ‘Getting Organized’: The Evolution of Community Interpreting. *Interpreting. International Journal of Research and Practice in Interpreting*,*4* (1), 125–140. DOI: 10.1075/intp.4.1.11poc.

Pöchhacker, Franz (2000). Language Barriers in Vienna Hospitals. *Ethnicity & Health*,*5* (2), 113–119.

Pöchhacker, Franz (2001). Quality Assessment in Conference and Community Interpreting. *Meta*,*46* (2), 410–425. DOI: 10.7202/003847ar.

Pöchhacker, Franz (2022). *Introducing Interpreting Studies*. London: Routledge. 3rd edn. DOI: 10.4324/9781003186472.

Putera, Lalu Jaswadi & Sujana, I. Made (2024). A comparative analysis of accuracy between Google Translate and Bard in translating abstracts of scientific journals. *Didaktik: Jurnal Ilmiah PGSD STKIP Subang*,*10* (1), 35–44.

Reckitt (2026). *The science of ibuprofen | Nurofen AU*. https://www.nurofen.com.au/pain-advice/about-nurofen/what-is-ibuprofen/ (Accessed: 24.3.2026).

Roy, Cynthia B. (1992). A Sociolinguistic Analysis of the Interpreter’s Role in Simultaneous Talk in a Face-to-Face Interpreted Dialogue. *Sign Language Studies*,*74*, 21–61.

Sabrina, Sabrina; Fadhillah, Mina; Chawdri, Flora Oktaviani; Safara, Cut & Juniarti, Lara (2025). A Contrastive Analysis of DeepL Translation vs. Google Translate’s Performance in Rendering Academic Texts: Insights from EFL Learners. *Jurnal Serambi Ilmu*,*26* (1), 73–82. DOI: 10.32672/jsi.v26i1.2502.

Saldanha, Gabriela (2014). *Research methodologies in translation studies*. London: Routledge. DOI: 10.4324/9781315760100.

Schegloff, Emanuel A.; Jefferson, Gail & Sacks, Harvey (1977). The Preference for Self-Correction in the Organization of Repair in Conversation. *Language*,*53* (2), 361. DOI: 10.2307/413107.

Seleskovitch, Danica (1968). *L’ interprète dans les conférences internationales : problèmes de langage et de communication*. Paris: Lettres Modernes.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*,*28* (3), 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

Sin-wai, Chan (2023). The Development of Translation Technology: 1967–2023. In: Sin-wai, Chan (Ed.) *Routledge Encyclopedia of Translation Technology*. London: Routledge. 2nd edn. 3–41. DOI: 10.4324/9781003168348.

Sourcenext Corporation (2025). *Pocketalk*. https://www.pocketalk.com/ (Accessed: 21.8.2025).

SPD; Bündnis 90/Die Grünen; & FDP (2021). *Mehr Fortschritt wagen. Koalitionsvertrag 2021–2025 zwischen der Sozialdemokratischen Partei Deutschlands (SPD), Bündnis 90/Die Grünen und den Freien Demokraten (FDP)*. Berlin: SPD, Bündnis 90/Die Grünen, FDP.

Stalnaker, Robert (1978). Assertion. In: Cole, Peter (Ed.) *Syntax and Semantics*. New York: Academic Press. Vol. 9. 315–332. https://archive.nytud.hu/program/szuperkurzus2006/Stalnaker_78.pdf.

Stewart, Moira A. (1995). Effective physician-patient communication and health outcomes: a review. *Canadian Medical Association Journal (CMAJ)*,*152* (9), 1423–1433.

Tan, Lee (2023). Speech Translation. In: *Routledge Encyclopedia of Translation Technology*. London: Routledge. 2nd edn. 726–737. DOI: 10.4324/9781003168348.

Thonon, Frédérique; Perrot, Swati; Yergolkar, Abhijna Vithal; Rousset-Torrente, Olivia; Griffith, James W.; Chassany, Olivier & Duracinsky, Martin (2021). Electronic Tools to Bridge the Language Gap in Health Care for People Who Have Migrated: Systematic Review. *Journal of Medical Internet Research*,*23* (5), 1–14. DOI: 10.2196/25131.

Turovsky, Barak (2016). *Found in translation: More accurate, fluent sentences in Google Translate*. https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/ (Accessed: 21.8.2025).

Turulski, Anna-Sofie (2024). *Ausländer in Wien nach Staatsangehörigkeiten 2024*. https://de.statista.com/statistik/daten/studie/886994/umfrage/auslaender-in-wien-nach-staatsangehoerigkeit/ (Accessed: 18.1.2025).

UNIVERSITAS Austria (2017). Berufs- und Ehrenordnung. UNIVERSITAS Austria. https://universitas.org/wp-content/uploads/Berufs-_und_Ehrenordung_2017_final-1.pdf.

Van Kolfschooten, Hannah; Goosen, Simone; Van Oirschot, Janneke; Schouten, Barbara; Vajda, Ildikó & Willems, Luna (2025). Legal, ethical, and policy challenges of artificial intelligence translation tools in healthcare. *Discover Public Health*,*22* (1), 904. DOI: 10.1186/s12982-025-01277-z.

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Łukasz & Polosukhin, Illia (2017). Attention is All you Need. *Neural Information Processing Systems*,.

Vieira, Lucas Nunes; O'Hagan, Minako & O'Sullivan, Carol (2021). Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society*,*24* (11), 1515–1532. DOI: 10.1080/1369118X.2020.1776370.

Wadensjö, Cecilia (1998). *Interpreting as Interaction*. London New York: Routledge.

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 Berlin Heidelberg. 3–21. DOI: 10.1007/978-3-662-04230-4_1.

Waibel, A.; Jain, A. N.; McNair, A. E.; Saito, H.; Hauptmann, A. G. & Tebelskis, J. (1991). *JANUS: a speech-to-speech translation system using connectionist and symbolic processing strategies*. In: *[Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing*. Toronto, Ont., Canada: IEEE. 793–796 vol.2. DOI: 10.1109/ICASSP.1991.150456.

Whyatt, Bogusława; Hatzidaki, Anna & French, Brian (2025). Participant profiling. In: Rojo López, Ana María, & Muñoz Martín, Ricardo (Eds) *Research Methods in Cognitive Translation and Interpreting Studies*. Amsterdam: John Benjamins Publishing Company. Vol. 10. 21–48. DOI: 10.1075/rmal.10.

Xiaojing, Bai & Shiwen, Yu (2023). Rule-Based Machine Translation. In: Sin-wai, Chan (Ed.) *Routledge Encyclopedia of Translation Technology*. London: Routledge. 2nd edn. 193–207. DOI: 10.4324/9781003168348.

Xueting, Liu & Chengze, Li (2023). Artificial Intelligence and Translation. In: Sin-wai, Chan (Ed.) *Routledge Encyclopedia of Translation Technology*. London: Routledge. 2nd edn. 280–301. DOI: 10.4324/9781003168348.

Yang, Jin & Lange, Elke (2003). Going live on the internet. In: Somers, Harold (Ed.) *Computers and Translation. A translator's guide.* Amsterdam: John Benjamins Publishing Company. Vol. 35. 191–210. DOI: 10.1075/btl.35.15yan.

Yang, Liu & Min, Zhang (2023). Statistical Machine Translation. In: Sin-wai (Ed.) *Routledge Encyclopedia of Translation Technology*. London: Routledge. 2nd edn. 208–218. DOI: 10.4324/9781003168348.

Zeng, Zhaohan & Liang, Zhibin (2024). Large Language Models are Good Translators. *Journal of Emerging Investigators*,*7*. DOI: 10.59720/24-020.

Appendix

Scenarios

Patient 1

Scenario 1A:

You are _____ (insert name), a 25-year-old student from Romania who came to Vienna for one semester in a student exchange. Although you do not speak any German, you have managed well thus far with English. However, for the last 3 days, you have been feeling worse and worse, vomiting multiple times a day, and suffering from diarrhoea, passing hardly any urine for two days, and feeling increasingly fatigued. This evening, you noticed your ankles were swollen, accompanied by a sudden dizziness, so you decided to go to the hospital.

Context:
Four days ago, you went on a hike with some new friends you met at the dorm. It was very hot outside (36 degrees C), and you only brought a 0.5L water bottle with you, assuming you could refill it along the way. You could not manage to refill your water, but when you got back to the dorm, you drank some beers with your friends. You do not usually smoke, but that night, you smoked 3 cigarettes with them because you were all having a great time together.
The next morning, you felt very hungover, had a bad headache and started to vomit. Even after you took ibuprofen, the vomiting did not stop. You noticed that your urine was dark brown, and you also started having diarrhoea. Since then, you have not urinated, and you have felt increasingly tired. This evening, you noticed that your ankles were very swollen, so you decided it was time to go to the hospital.
At the emergency room, you waited 20 minutes for a doctor to call you in, and when she did, you were barely able to walk to the consultation room and had to hold onto the wall, since you suddenly got dizzy.

Task:
Communicate to the best of your ability with the doctor to find out what is wrong and what you have to do next. The consultation will last approximately 10 minutes.

Doctor 1A

You are a doctor on a regular night shift in the emergency room. At 22:00, a male patient arrives at the hospital. You have been told that the patient only speaks Romanian. When you open the door to call the patient in, you notice that he has to support himself on the wall after standing up from the chair, indicating dizziness. Since it is late in the evening, and the patient needs immediate attention, the option to call a professional interpreter is not available, so you decide to use ChatGPT on the phone.

Task:
Take the patient's medical history and communicate the next steps to him using the app on the phone. You have 10 min.

Scenario 1B:

You are _____ (insert name), a 27-year-old post-doc from Romania who come to Vienna for a conference, and you only speak Romanian. After a long train journey, you started to experience worsening upper abdominal pain. It quickly turned into a stabbing pain radiating straight through to your back, making you extremely uncomfortable and unable to sit or lie flat. After vomiting three times and taking over-the-counter pain medication without effect, you decided to go to the hospital.

Context:
Over the past week, you have been celebrating a lot after your most recent work has been published, including several evenings of heavy alcohol consumption. After not getting enough sleep for the past few days, you took the train to Vienna, and upon arriving, you started feeling discomfort in the upper abdominal area. Around lunchtime, you did not have any appetite, even though your last meal was at midnight, when you ate the kebab, you bought at the train station before departure.
You were diagnosed with severe hypertriglyceridemia three years ago, but you were not given any medication. You do not have a history of similar abdominal pain. You normally do not get travel sickness, but after lunch you started feeling nauseous, and the dull pain turned into a stabbing pain expanding to your back. You had three episodes of non-bloody, non-bilious vomiting, but had no diarrhoea, chest pain, shortness of breath, or urinary symptoms. Because the pain became unbearable, you decided to go to the hospital.

Task:
Communicate to the best of your ability with the doctor to find out what is wrong and what you have to do next. The consultation will last approximately 10 minutes.

Doctor 1B

You are a doctor on a regular shift in the emergency room. In the late afternoon, a male patient arrives at the hospital. You have been told that the patient only speaks Romanian. When you open the door to call the patient in, you notice that he is in agonising pain and holding his back. Luckily, the hospital has an interpreter for Romanian, and she is assisting the patient.

Task:
Take the patient's medical history and communicate the next steps to him with the help of the interpreter. You have 10 min.

Patient 2

Scenario 2A:

You are _____ (insert name), a 23-year-old male from Bucharest, who went on a road trip through Europe with his partner. Neither of you knows any other language apart from Romanian. You have been on a trip for the last three days, but on the 4th day, you started to feel pain around your belly button at around 11:00. The pain got progressively worse and moved to the lower right part of the abdomen. In the afternoon, the pain got so bad that you could not stand up straight or walk anymore. You are exhausted and nauseous and decide to go to the hospital.

Context:
You started your trip four days ago from Bucharest and stopped for the last three nights along the road in hostels to sleep. This morning, you started driving from Budapest to Vienna and only stopped for coffee before reaching the Austrian border. Around 11:00, you started feeling a dull pain around the belly button, but you thought it might be because you drank coffee on an empty stomach, so you did not think much of it.
After arriving in Vienna, you checked in at the hotel. By the time you left the hotel to take a walk through the city, the pain had gotten sharper and had moved to the lower right part of your abdomen. Walking became suddenly exhausting. You felt nauseated, sweaty, and lightheaded.

You returned to your hostel, hoping the pain would pass, but an hour later, it had worsened so much that you could not stand up straight. You ask your partner to call a taxi and take you to the hospital.

Task:
Communicate to the best of your ability with the doctor to find out what is wrong and what you have to do next. The consultation will last approximately 10 minutes.

Doctor 2A

You are a doctor on a regular shift in the emergency room. In the late afternoon, a male patient arrives at the hospital. You have been told that the patient only speaks Romanian. When you open the door to call the patient in, you notice that he is walking very stiffly, grimacing in pain and holding his abdomen. Luckily, the hospital has an interpreter for Romanian, and she is assisting the patient.

Task:
Take the patient's medical history and communicate the next steps to him with the help of the interpreter. You have 10 min.

Scenario 2B:

You are _____ (insert name), a 28-year-old Romanian PhD student in Vienna who only speaks Romanian and English. For the last five days, you have had a worsening cough, fever, and shortness of breath.

Context:
A few days ago, you went to the Christmas market with your peers, and it started raining. That evening, you developed a dry cough and mild sore throat, but you thought it was just a cold, and it often happens to you, since you are a smoker. Over the next two days, your cough became productive, bringing up yellow-green sputum, and you began experiencing sharp right-sided chest pain that worsened when you took deep breaths.
You tried taking over-the-counter cold medicine, but your symptoms continued to worsen. Last night, you woke up several times because you felt short of breath, especially when lying down. You also had a fever and noticed you were breathing faster than usual. Even though you spent

yesterday in bed, mundane tasks like going to the bathroom, going out for a smoke and making soup felt exhausting, and you became lightheaded when standing up.
On the fifth day, after seeing early in the morning that the symptoms persisted and you felt weaker and weaker, you decided to go to the hospital.

Task:
Communicate to the best of your ability with the doctor to find out what is wrong and what you have to do next. The consultation will last approximately 10 minutes.

Doctor 2B

You are a doctor on a regular shift at the emergency department. At 7:30, a male patient arrives at the hospital. You have been told that the patient only speaks Romanian. When you open the door to call the patient in, you notice his flushed cheeks and hear him coughing repeatedly. Since it is early in the morning, and the patient needs immediate attention, the option to call a professional interpreter is not available, so you decide to use ChatGPT on the phone.

Task:
Take the patient's medical history and communicate the next steps to him using the app on the phone. You have 10 min.

Interview Questions for Patients

Section 1: General communication experience and empathy

- Do you have prior experience with interpreters or with the interpreter function of ChatGPT?
- How would you describe your overall communication experience during these consultations?
- How well do you think the healthcare provider understood your health complaints and concerns?
- Did you feel that you understood the doctor's questions, explanations, and instructions?
- Were there any moments where you were unsure about what the doctor meant?
- Were there any parts of the consultation you found confusing?

Section 2: S2S vs. human interpreting

- Do you think one interpreting mode was more helpful than the other? If so, which one?
- Which interpreting mode transported your message best? What contributed to that?
- How did the interpretation mode affect the conversational and emotional flow of the consultation?
- Did you change the way you usually speak to facilitate communication with the interpreter or the S2S system?
- Did the interpreter or the S2S system help you feel heard and understood?
- How empathetic did the healthcare provider seem during each consultation? How important is that to you?

Section 3: Perception of accuracy and errors

- Was there any unexpected reaction from the doctor that indicated a discrepancy between what you said and what was understood through the interpretation?
- Did you notice any information being left out, changed, or added by the interpreter or S2S system?
- Did you feel that tone, urgency, or emotional content was correctly conveyed?
- Did you feel there was a difference in how medical terminology was used in each mode?
- How did accuracy errors, if any, affect the flow of the consultation?

Section 4: Closing questions

- If you had to choose between the two interpreting modes as a patient, which would you prefer, and why?
- Is there anything else you noticed that you think is important to mention?

Interview Questions for the Healthcare Provider

Section 1: General communication experience

- Do you have prior experience with interpreters or with the interpreter function of ChatGPT?
- How would you describe your overall communication experience during these consultations?
- How well did you understand the patients' health complaints and concerns?
- Did you feel that the patients understood your questions, explanations, and instructions?
- Were there any moments where you were unsure about what the patients meant?
- Were there any parts of the consultation you found confusing?

Section 2: S2S vs. human interpreting

- Do you think one interpreting mode was more helpful than the other? If so, which one?
- Which interpreting mode transported your message best? What contributed to that?
- How did the interpretation mode affect the conversational and emotional flow of the consultation?
- Did you change the way you usually speak to facilitate communication with the interpreter or the S2S system?
- Did the interpreter or the S2S system enhance or hinder your ability to connect with the patients?
- Do you think the patients perceived your empathy through the interpreter or the S2S system? If so, how? How important is that to you

Section 3: Perception of accuracy and errors

- Was there any unexpected reaction from any patient that indicated a discrepancy between what you said and what was understood through the interpretation?

- Did you notice any information being left out, changed, or added by the interpreter or S2S system?
- Did you feel that tone, urgency, or emotional content was correctly conveyed?
- Did you feel there was a difference in how medical terminology was used in each mode?
- How did accuracy errors, if any, affect the flow of the consultation?

Section 4: Closing questions

- If you had to choose between the two interpreting modes as a doctor, which would you prefer, and why?
- Is there anything else you noticed that you think is important to mention?

Analysis Tables

C1

Segment nr.	Speaker	Utterance	Analysis		
			Hamai et al. (2024)	MQM	Clark & Schaefer (1989)
1.	A1.1	Guten Abend, grüß Gott, Dr. Heindrich ist mein Name.	-	-	-
2.	D1.1	Bună ziua, bună seara, doctor Heindrich e numele meu.	p	-	-
3.	A1.1	Was führt Sie denn zu uns?	-	-	-
4.	D1.1	De ce ați venit la noi?	p	-	-
5.	P1.1	Bună, în primul rând, mă numesc George, eram cu mașina, mergeam cu prietena, sunt pe drum de trei zile, și într-o excursie prin Europa, și în a patra zi azi dimineață, când am plecat din Budapesta, am băut doar o cafea la granița cu Austria. Și probabil din cauza asta, dar nu sunt sigur de ce, a început să mai ia o durere de burtă super nasoală (**D1.1**: Aham). Progresiv durerea s-a amplificat. Am ajuns în Viena, ne-am plimbat, dar pe cum a trecut timpul, m-a durut burta și mai mult. Și la un moment dat, pur și simplu simțeam că nu mai puteam să merg (**D1.1**: Aham). Și eram mult prea obosit. Am decis să mă duc înapoi la pensiunea unde eram cazați. Și am crezut că o să-mi revin, dar nu mi-am revenit. În continuare mă simt foarte prost. Am avut stări de greață. Am transpirat foarte mult. M-am pus în pat. N-am mai putut sta în picioare. Și acuma am rugat-o pe prietena mea să mă ducă la spital. De asta sunt aici.	-	-	IRNC of P1.1 ACK of D1.1
6.	D1.1	Also, erstens möchte ich sagen, ich heiße George. Ich bin seit mehreren Tagen unterwegs, seit drei Tagen, mit meiner Freundin. Wir wollten einen Ausflug durch Europa machen. [Jetzt sind wir] Am vierten Tag sind wir aus Budapest losgefahren und am Morgen habe ich am nüchternen Magen nur einen Kaffee getrunken bei der Grenze zu Österreich. Ab dem Moment habe	s+	-	-

		ich begonnen Magenbeschwerden zu haben, Magenschmerzen, die immer schlimmer geworden sind. Ich weiß nicht [warum das so war], warum das so ist. Ich nehme mal an, dass es wegen dem Kaffee ist. Letztendlich sind wir nach Wien gekommen. Wir sind ein bisschen spazieren gegangen, in der Hoffnung, dass, wenn die Zeit vergeht, es etwas besser wird, aber ganz im Gegenteil. Die Schmerzen sind schlimmer geworden. Ich konnte nicht mehr weitergehen, ich fühlte mich sehr müde, also wollten wir zurück zu unserer Unterkunft in der Annahme, dass es mir bald wieder besser geht, aber es war nicht der Fall. Ich hab auch Übelkeit verspürt. Ich hab viel geschwitzt. Ich hab mich dann hinlegen müssen, weil ich nicht mehr stehen konnte, und letztendlich hab ich dann meine Freundin gebeten mich ins Krankenhaus zu bringen, deswegen bin ich jetzt bei Ihnen.			
7.	A1.1	Alles klar, okay. Hatten Sie das schon einmal oder ist das die erste Episode?	-	-	ACK of A1.1
8.	D1.1	Ați mai avut vreodată stări similare sau este prima dată (**P1.1**: Nu, nu) în viață când vi se întâmplă?	p	-	-
9.	P1.1	E prima oară.	-	-	IRNC
10.	D1.1	Nein, es ist das erste Mal.	p	-	-
11.	A1.1	Und die Schmerzen, auf einer Skala von eins bis zehn, wie würden Sie die einschätzen?	-	-	IRNC of A1.1
12.	D1.1	Cum ați estima durerile de la unu la zece?	p	-	-
13.	P1.1	Probabil un opt. În momentul actual.	-	-	IRNC of P1.1
14.	D1.1	Wahrscheinlich eher eine Acht (**A1.1**: Aham) im Moment.	p	a.a minor	ACK of A1.1
15.	A1.1	Hat er schon irgendwelche Medikamente eingenommen heute?	-	-	Pronoun switch
16.	D1.1	Ați luat medicamente astăzi deja?	s+	-	-
17.	P1.1	Nu, n-am luat nimic.	-	-	IRNC
18.	D1.1	Nein, keine.	p	-	-
19.	A1.1	Gibt's eine Dauermedikation?	-	-	-
20.	D1.1	Există medicație de durată pe care o luați continuu?	p	-	-
21.	P1.1	Nu.	-	-	IRNC
22.	D1.1	Auch keine.	p	-	-

23.	A1.1	Und Vorerkrankungen?	-	-	-
24.	D1.1	Aveți cumva alte boli cronice?	s-	a.mt major	-
25.	P1.1	Nu, nimic.	-	-	IRNC
26.	D1.1	Nein, keine.	p	-	-
27.	A1.1	Nein, okay. Danke. Der Schmerz: Wenn Sie den beschreiben würden, ist es eher ein Stechen, ein Ziehen, ein Brennen, ein wellenartiger Schmerz?	-	-	ACK of A1.1 IRNC of A1.1 DISP of A1.1
28.	D1.1	Cum ați descrie durerea? Este o înțepătură? Este mai degrabă usturime? Este crampe sau este în valuri?	s	a.mt minor	-
29.	P1.1	E, hmm, cum să zic, mai ascuțită durerea în partea din dreapta jos a abdomenului ca niște cuțite, să zicem.	-	-	IRNC of P1.1
30.	D1.1	Es ist eher stechend, rechts unten, im unteren Bereich des Oberkörpers und es fühlt sich an wie Messerstiche.	p	-	-
31.	A1.1	Aham. Hat er erbrochen auch?			ACK of A1.1 Pronoun switch
32.	D1.1	Ați și vomitat?	s+	-	-
33.	P1.1	Nu, nu am vomitat, am avut stări de greață, dar n-am vomitat.	-	-	IRNC of P1.1
34.	D1.1	Nein, aber Übelkeit habe ich gespürt.	p	-	-
35.	A1.1	Gut. Das ist jetzt im rechten Oberbauch oder im Unterbauch?	-	-	ACK of A1.1
36.	D1.1	Abdomen dreapta sus sau jos?	-	-	-
37.	P1.1	Dreapta jos.	-	-	IRNC of P1.1
38.	D1.1	Rechts unten.	p	-	-
39.	A1.1	Rechter Unterbauch, okay, gut. Wie schaut's aus mit dem Stuhlgang? Irgendwelcher in Richtung Durchfall gehabt auch heute oder die letzten Tage?	-	-	DEMO of A1.1 ACK of A1.1
40.	D1.1	Cum e cu scaunul? Ați avut diaree astăzi sau în ultimele zile?	o+	s.as minor	-
41.	P1.1	Nu, nu am avut diaree.	-	-	IRNC of P1.1
42.	D1.1	Nein, kein Durchfall.	p	-	-
43.	A1.1	Okay, und außer dem Kaffee haben Sie heute noch nichts zu sich genommen?	-	-	ACK of A1.1
44.	D1.1	În afară de cafea, [nu ați luat] nu ați mâncat astăzi nimic?	p	-	-

45.	P1.1	Nu, nimic. (**D1.1**: Nein) Doar am băut cafeaua. Atât.	p	-	IRNC of P1.1
46.	D1.1	Nur den Kaffee (**A1.1**: Okay), über den ich gesagt hab.	s+	-	ACK of A1.1
47.	A1.1	Gut, okay. Das hamma. Gestern irgendwas gegessen?	-	-	ACK of A1.1
48.	D1.1	Ieri ați mâncat ceva? (**A1.1**: Was Außergewöhnliches irgendwie?) [Ceva ce nu era] Ceva neobișnuit?	p	-	-
49.	P1.1	Nu, nimic anormal, adică doar șnițel la un restaurant, adică nimic special.	-	-	IRNC of P1.1
50.	D1.1	Nichts Besonderes. (**A1.1**: Aham) Nur ein Schnitzel in einem Lokal.	p	-	ACK of A1.1
51.	A1.1	Gut, und die Freundin ist gesund?	-	-	ACK of A1.1
52.	D1.1	Prietena e sănătoasă? Ei îi merge bine?	s+	-	
53.	P1.1	Da, (**D1.1**: Ja, sie ist gesund) ea e okay.	p	-	IRNC of P1.1
54.	A1.1	Gut, okay. Schwitzen haben Sie gesagt…Fieber auch gehabt?	-	-	ACK of A1.1 DEMO of A1.1
55.	D1.1	Ați spus că ați transpirat. Ați avut și febră? (**A1.1**: Haben Sie es gemessen?) Sau ați măsurat febra?	p	-	-
56.	P1.1	Nu am măsurat febra (**D1.1**: Okay). Doar am început să transpir foarte mult, dar mi-a fost cald. (**D1.1**: V-a fost cald?)	v+	-	ACK of D1.1 IRNC
57.	D1.1	Ich hab's nicht gemessen, ich hab viel geschwitzt und es war mir heiß.	p	-	-
58.	A1.1	Okay. Schüttelfrost?	-	-	ACK of A1.1
59.	D1.1	Ați avut și frisoane?	a+	a.a minor	-
60.	P1.1	Da, (**D1.1**: Ja) am avut.	p	-	IRNC of P1.1
61.	A1.1	Okay. Haben Sie eine Allergie?	-	-	ACK of A1.1
62.	D1.1	Aveți cumva vreo alergie?	p	-	-
63.	P1.1	Nu, niciuna.	-	-	IRNC of P1.1
64.	D1.1	Nein, keine.	p	-	-
65.	A1.1	Gut. Also, ich hab jetzt mal so ein grobes Bild. Ja? [Wir würden] im nächsten Schritt würden Ihnen die Schwestern einmal Blut abnehmen und einen Venflon setzen, also einen venösen Zugang.	-	-	ACK of A1.1

66.	D1.1	Deci (**A1.1**: Ahm, genau), în prima fază am o imagine de ansamblu. În următorul pas asistentele vă vor lua sânge și vă vor pune o branulă. Da? (**P1.1**: Okay; **A1.1**: Ja?) Intravenoasă.	p	-	ACK of P1.1 ACK of A1.1
67.	A1.1	Ich mach dann eine körperliche Untersuchung, ja? Und Sie kriegen einmal von uns erstens ein Schmerzmittel, ja? Dass diese Beschwerden besser werden, ja? Und wir werden Sie auch zum Röntgen schicken, ja? Einmal vom Oberkörper und auch vom Bauch, also so eine Durchleuchtung, um zu sehen, ob da irgendwas los ist, und eventuell auch noch einen Ultraschall machen. Das könn'ma hier in der Ambulanz machen, den Ultraschall, und dann sehn'ma mal was da los ist, ja?	-	-	IRNC
68.	D1.1	Apoi voi face și un control, vă voi palpa, veți primi și un analgezic, să nu mai aveți durerile pe care le aveți acum. Avem de gând să facem și o radiografie de la abdomenul superior și zona burții. Pentru a vedea despre ce este vorba, putem face și un ecograf, pe care îl facem aici în clinica de zi, avem posibilitatea aceasta, și apoi vedem despre ce e vorba.	s+	-	-
69.	P1.1	Okay.	-	-	ACK of P1.1
70.	A1.1	Beim Urinieren noch irgendwelche Auffälligkeiten?	-	-	-
71.	D1.1	Când urinați, ați observat ceva neobișnuit?	p	-	-
72.	P1.1	Nu.	-	-	IRNC
73.	D1.1	No.	p	-	-
74.	A1.1	Passt, dann ist es in Ordnung. Okay, gut. Ich schau noch, ob ich irgendwas vergessen habe.	-	-	ACK of A1.1
75.	D1.1	Verific să văd dacă mai am nevoie de ceva.	p	-	-
76.	A1.1	Alkohol getrunken, (**D1.1**: Ați consumat alcool în ultimele zile?) in den letzten Tagen?	p	-	-
77.	P1.1	Da.	-	-	IRNC
78.	D1.1	Ja.	p	-	-
79.	A1.1	Und in welchem Ausmaß?	-	-	IRNC of A1.1

80.	D1.1	În ce măsură cantitativ?	a	a.a minor	-
81.	P1.1	Aproximativ patru beri în fiecare zi.	-	-	IRNC of P1.1
82.	D1.1	Um die vier Bier pro Tag.	p	-	-
83.	A1.1	Aham. Okay, gut. Gib's eine Kontaktperson, oder von der Freundin die Telefonnummer vielleicht, dass ma uns noch aufschreiben.	-	-	ACK of A1.1
84.	D1.1	Aveți o persoană de contact, poate ne dați numărat telefon al prietenei dumneavoastră?	p	-	-
85.	P1.1	Sigur, sigur.	-	-	IRNC
86.	D1.1	Mit Sicherheit.	p	-	-
87.	A1.1	Aham. Gut, das mach'ma dann später, so. Gut! Dann hab ich erstmal alles. Haben Sie jetzt noch akut Fragen, bevor wir starten mit den weiteren Untersuchungen?	-	-	ACK of A1.1
88.	D1.1	Numărul de telefon ni-l spuneți mai tarziu (**P1.1**: Aham). Din punctul meu de vedere acestea au fost toate întrebările, mai aveți dumneavoastră, întrebări? (**P1.1:** Nu) Înainte de să începem cu investigațiile.	p	-	ACK of P1.1
89.	P1.1	Nu, putem să începem.	-	-	IRNC
90.	D1.1	Nein, (**A1.1**: Alles klar) wir können beginnen.	p	-	-
91.	A1.1	Gut. Dann fang'ma an jetzt mit den Untersuchungen.	-	-	ACK of A1.1
92.	D1.1	Bun, începem.	o+	-	-
93.	P1.1	Super.	-	-	-

C2

Segment nr.	Speaker	Utterance	Analysis		
			Hamai et al. (2024)	MQM	Clark & Schaefer (1989)
1.	A1.2	Transcript Okay, guten Tag, Dr. Händrich ist mein Name. Was führt Sie zu uns in die Notfallambulanz? Voice Okay, guten Tag, Dr. Heindrich ist mein Name. Was führt Sie zu uns in die Notfallambulanz?	-	-	-

2.	ChatGPT	Okay, bună ziua, numele meu este doctor Händi, ce vă aduce la noi în camera de urgență?	s	a.mt minor	-
3.	P1.2	Păi mă simt foarte prost. În ultimele 5 zile am început să am o tuse productivă și am avut și febră și dificultăți de respirație și nu știu de ce se întâmplă. Mă simt foarte prost.	-	-	IRNC
4.	ChatGPT	Also, mir geht es wirklich schlecht, seit ungefähr fünf Tagen habe ich einen produktiven Husten, ich hatte auch Fieber und Atembeschwerden, und ich weiß nicht, was los ist, ich fühle mich wirklich schlecht.	p	-	-
5.	A1.2	Haben Sie Atemnot?	-	-	DEMO
6.	ChatGPT	Aveți dificultăți de respirație?	p	s.as minor	-
7.	P1.2	Transcript - Voice Da.	o	a.o minor	IRNC
8.	A1.2	Nur bei Ruhe oder auch bei Belastung.	-	-	IRNC
9.	ChatGPT	Și doar în repaus sau și la efort?	p	-	-
10.	P1.2	Transcript Țin repauz și la efort! Voice Și în repaus și la effort.	-	-	IRNC
11.	ChatGPT	Sowohl in Ruhe als auch bei Belastung.	p	-	-
12.	A1.2	Und der Husten war mit Auswurf. Wie schaut denn das aus, wenn Sie aushusten müssen? Welche Farbe hat denn das?	-	-	DEMO of A1.2 IRNC of A1.2
13.	ChatGPT	Și tusea a fost cu expectorație, cum arată atunci când trebuie să expectorați, ce culoare are aceasta?	p	-	-
14.	P1.2	Da, a fost cu expectorație, a fost de culoare galben-verzuie.	-	-	IRNC
15.	ChatGPT	Ja, es war eine Expektoration, sie hatte eine gelblich-grüne Farbe.	-	-	-
16.	A1.2	Und bis wie viel Grad ist das Fieber hinaufgegangen?	-	-	IRNC
17.	ChatGPT	Și până la câte grade a ajuns febra?	p	-	-
18.	P1.2	Transcript - Voice Mmm (**ChatGPT:** Alles klar) 38 cu 6.	a	a.a minor	IRNC

19.	P1.2	Transcript Trej šopku šase. Voice -	-	-	-
20.	ChatGPT	Transcript 36,6 Grad. Voice 63,6 Grad.	s-	a.mt critical	-
21.	A1.2	Können Sie es noch mal wiederholen, wie hoch das Fieber war?	-	-	-
22.	ChatGPT	Poți să repeți din nou cât de mare a fost febra?	s	s.lr minor	-
23.	P1.2	Transcript 38. 6 grade Celsius. Voice 38 virgulă 6 grade Celsius.	-	-	IRNC
24.	ChatGPT	Transcript 36,6 Grad Celsius. Voice 63 Komma 6 Grad Celsius.	s-	a.mt critical	-
25.	A1.2	Transcript - Voice Gibt es Vorerkrankungen bei Ihnen?	o-	a.o major	-
26.	A1.2	Gibt es Vorerkrankungen?	-	-	-
27.	ChatGPT	Aveți antecedente medicale?	p	-	-
28.	P1.2	Transcript Nu am incidente medicale, doar că eu sunt și fumător. Inițial credeam că aceste simptome sunt doar de la o răceală sau pentru că sunt fumător, că mai tușesc, dar s-au agravat foarte mult și de asta am venit la spital. Voice Nu, nu am antecedente medicale, doar că eu sunt și fumător. Inițial credeam că aceste simptome sunt doar de la o răceală sau pentru că sunt fumător, că mai tușesc, dar s-au agravat foarte mult și de asta am venit la spital.	-	-	IRNC
29.	ChatGPT	Ich habe keine Vorerkrankungen , aber ich bin auch Raucher, und anfangs dachte ich, dass diese Symptome einfach nur von einer Erkältung oder vom Rauchen selbst kommen, weil ich öfter huste, aber es hat sich dann sehr verschlechtert, und	p	-	-

		deshalb bin ich ins Krankenhaus gekommen.			
30.	A1.2	Transcript Wie viele Zigaretten rauchen Sie denn am Tag? Voice Wie viele rauchen Sie denn am Tag? Wie viele Zigaretten?	o-	a.o major	IRNC
31.	A1.2	Wie viele Zigaretten pro Tag rauchen Sie?	-	-	-
32.	ChatGPT	Câte țigări fumați pe zi?	p	-	-
33.	P1.2	Jumate de pachet.	-	-	IRNC
34.	ChatGPT	Eine halbe Schachtel.	p	-	-
35.	A1.2	Und seit wie vielen Jahren?	-	-	IRNC
36.	ChatGPT	Atunci de câți ani?	p	s.as minor	-
37.	P1.2	Transcript Depeche Pane. Voice De paișpe ani.	-	-	IRNC
38.	ChatGPT	Zum Beispiel seit…	s-	a.mt major	-
39.	A1.2	Seit wie vielen Jahren sagen Sie das nochmal? Das hat es, glaube ich, nicht verstanden.	-	-	-
40.	ChatGPT	De câți ani, spuneți-mi din nou, cred că nu am înțeles.	s	a.mt minor	-
41.	P1.2	Transcript De 14 ani, cum es a început, nu mult, dar de 14 ani in total. Voice De 14 ani, cum am zis. La început, nu mult, dar de 14 ani în total.	-	-	IRNC
42.	ChatGPT	Seit 14 Jahren rauche ich, es hat nicht vor langer Zeit begonnen, aber seit 14 Jahren insgesamt.	s-	a.mt major	-
43.	A1.2	Nehmen Sie regelmäßig Medikamente ein.	-	-	-
44.	ChatGPT	Luați medicamente în mod regulat?	p	-	-
45.	P1.2	De obicei nu, nu iau medicamente, dar am luat niște medicamente acum de când am avut niște simptome, doar pentru răceală, fără prescripție, dar nu m-au ajutat deloc.	-	-	IRNC
46.	ChatGPT	Normalerweise nicht, ich nehme keine Medikamente, aber ich habe jetzt einige Medikamente eingenommen, seit ich diese Symptome hatte, nur für die	p	-	-

		Erkältung, ohne Rezept, aber sie haben mir nicht wirklich geholfen.			
47.	A1.2	Haben Sie da ein Foto oder eine Liste von den Medikamenten mit?	-	-	IRNC
48.	ChatGPT	Aveți vreo fotografie sau o listă cu medicamentele respective?	p	-	-
49.	P1.2	Nu, dar știu că am luat paracetamol și Nurofen și vitamina C.	-	-	IRNC
50.	ChatGPT	Nein, aber ich weiß, dass ich Paracetamol und Ibuprofen sowie Vitamin C eingenommen habe.	s+	-	-
51.	A1.2	Besteht auch ein Thoraxschmerz?	-	-	-
52.	ChatGPT	Există și o durere toracică?	p	-	-
53.	P1.2	Transcript Nu uitați să dați like, să lăsați un comentariu și să distribuiți acest material video pe alte rețele sociale. Voice Mmm, da, adică mă doare în partea dreaptă a pieptului și mă doare și mai tare când respir adânc. Se agravează durerea când respir adânc.	-	-	IRNC
54.	ChatGPT	Ja, es schmerzt auf der rechten Seite der Brust, und der Schmerz wird noch stärker, wenn ich tief einatme, der Schmerz verschlimmert sich also beim tiefen Atmen.	p	-	-
55.	A1.2	Alles klar, okay. Wir werden jetzt eine Diagnostik machen. Ja, also zuerst mal werden Sie einen Zugang bekommen. Wir machen eine Blutabnahme und auch eine Blutabnahme aus der Arterie, um zu schauen, ob Sie momentan Sauerstoff brauchen.	-	-	-
56.	ChatGPT	În regulă, bine, acum vom efectua o diagnosticare, da, deci, mai întâi, vă vom pune un acces, vom face o recoltare de sânge și, de asemenea, o recoltare de sânge din arteră, pentru a verifica dacă aveți nevoie în prezent de oxigen.	p	s.as minor	-
57.	P1.2	Transcript - Voice Okay, am înțeles.	o	a.o minor	-
58.	A1.2	Transcript Okay, haben wir alles.Und dann schicken wir sie zum Röntgen, um zu schauen, was in der Lunge los ist.	-	-	-

		Das könnte zum Beispiel eine Lungenentzündung sein. Und ich werde sie abhören und generell untersuchen. Voice Und dann schicken wir sie zum Röntgen, ja? Um zu schauen, was in der Lunge los ist. Das könnte zum Beispiel eine Lungenentzündung sein, ja? Und ich werde sie abhören und generell untersuchen.			
59.	ChatGPT	Și apoi vă vom trimite la radiografie, da, pentru a vedea ce se întâmplă în plămâni, ar putea fi, de exemplu, o pneumonie, da, și vă voi asculta cu stetoscopul și vă voi examina în general.	p	s.as minor	-
60.	P1.2	Transcript - Voice Okay, super.	o	a.o minor	ACK of P1.2
61.	A1.2	Transcript Okay, super. Und je nachdem, was da herauskommt, kriegen Sie dann die entsprechende Therapie. Und eine letzte Frage habe ich noch, bevor wir starten. Haben Sie einen Lungenfacharzt? Voice Und je nachdem, was da herauskommt, kriegen Sie dann die entsprechende Therapie. Und eine letzte Frage habe ich noch, bevor wir starten. Haben Sie einen Lungenfacharzt?	-	-	-
62.	ChatGPT	Transcript Bine, minunat! Și, în funcție de ceea ce va rezulta, veți primi apoi terapia corespunzătoare, iar o ultimă întrebare înainte să începem: Voice Bine, minunat! Și, în funcție de ceea ce (**A1.2**: Ha…)	a o-	a.a minor a.o major	-
63.	A1.2	Transcript Haben Sie einen Lungenfarkt? Voice Haben Sie einen Lungenfacharzt?	-	-	-
64.	ChatGPT	Aveți un specialist pneumolog?	p	-	-
65.	P1.2	Transcript No. No. No.	-	-	IRNC

		Voice Nu, nu am.			
66.	ChatGPT	Nein, nein.	p	-	-
67.	A1.2	Alles klar, haben Sie jetzt noch Fragen, bevor wir starten?	-	-	ACK
68.	ChatGPT	În regulă, mai aveți acum întrebări înainte să începem?	p	-	-
69.	P1.2	Transcript - Voice Nu, putem începe.	-	-	IRNC

C3

Segment nr.	Speaker	Utterance	Analysis		
			Hamai et al. (2024)	MQM	Clark & Schaefer (1989)
1.	A2.1	Grüß Gott, Dr. Heindrich ist mein Name. Was führt Sie zu uns in die Notaufnahme?	-	-	-
2.	D2.1	Bună ziua, dr. Heindrich mă numesc. De ce veniți la noi aici la urgențe?	p	s.as minor	-
3.	P2.1	Bună ziua. Am venit, deoarece în ultimele zile am avut probleme foarte mari cu stomacul. Am avut diaree și vărsături. De asemenea, nu am mai fost la baie să urinez de două zile și nu știu ce se întâmplă. De asemenea, gleznele mele s-au umflat foarte mult acuma. Și da, mă simt foarte prost.	-	-	IRNC
4.	D2.1	Ja, ich bin gekommen, weil in den letzten Tagen hatte ich viele Probleme mit dem Magen hatte, viele Magenbeschwerden. Ich hatte Durchfall und Erbrechen. [Ich bin seit zwei Tagen nicht mehr] Ich konnte seit zwei Tagen nicht mehr urinieren. Ich weiß nicht ganz genau was passiert ist. Ich hab bemerkt, dass [meine] mein Sprunggelenk sehr angeschwollen ist.	o	a.o minor	-
5.	A2.1	Aham. Und die ganzen Beschwerden sind seit zwei Tagen?	-	-	ACK
6.	D2.1	Toate aceste probleme le aveți de două zile?	p	-	-
7.	P2.1	Da, de două zile. Vomit regulat și de asemenea și am diaree.	-	-	IRNC

8.	D2.1	Aham. Ja, seit zwei Tagen (**A2.1**: Aham) erbreche ich regelmäßig und habe Durchfall.	p	-	ACK of D2.1 ACK of A2.1
9.	A2.1	Okay. Das Erbrechen, ist das immer nach dem Essen?	-	-	ACK IRNC
10.	D2.1	Vărsăturile vin după mâncare?	p	-	-
11.	P2.1	Nu, nu neapărat, ci chiar dacă nu mănânc.	-	-	IRNC
12.	D2.1	Nein, nicht unbedingt, die kommen, auch wenn ich nicht esse.	p	-	-
13.	A2.1	Gibt's bei Ihnen Vorerkrankungen?	-	-	-
14.	D2.1	Există alte boli pe care le aveți? Antecedente?	p	-	-
15.	P2.1	Nu. Nu am avut alte probleme în trecut.	-	-	IRNC
16.	D2.1	Nein, ich hatte keine anderen Beschwerden in der Vergangenheit.	p	-	-
17.	A2.1	Operationen schon gehabt im Magen-Darm-Trakt?	-	-	-
18.	D2.1	Ați avut cumva operații, intervenții chirurgicale în zona gastrointestinală?	p	-	-
19.	P2.1	Nu, nu am fost operat.	-	-	IRNC
20.	D2.1	Nein, keine (**A2.1:** Okay).	p	-	ACK of A2.1
21.	A2.1	Nehmen Sie regelmäßig Medikamente ein?	-	-	-
22.	D2.1	Luați medicamente mod regulat?	p	-	-
23.	P2.1	Nu, dar am luat în ultimele zile ibuprofen, deoarece am avut vărsături, dar nu a avut niciun efect.	-	-	IRNC
24.	D2.1	Nein, aber in den letzten Tagen habe ich Ibuprofen angenommen, gegen dem Erbrechen, aber das hatte gar keine Wirkung gehabt.	p	-	-
25.	A2.1	Das war das einzige Medikament. Sonst nix?	-	-	-
26.	D2.1	Singurul medicament pe care l-ați luat? În rest nimic?	p	-	-
27.	P2.1	Da.	-	-	IRNC
28.	D2.1	Ja.	p	-	-
29.	A2.1	Ja, okay. Waren Sie damit beim Hausarzt oder haben Sie das einfach so in der Apotheke, oder zu Hause gehabt?	-	-	ACK
30.	D2.1	Ați fost la medicul de familie pentru acest medicament, sau l-ați avut acasă sau ați fost la farmacie și l-ați cumpărat pur și simplu?	p	-	-
31.	P2.1	Am avut acasă. Nu am fost la farmacie.	-	-	IRNC

32.	D2.1	Ich hatte noch zu Hause (**A2.1:** Zuhause, okay), ich musste nicht in die Apotheke gehen.	p	-	DISP of A2.1
33.	A2.1	Wie schaut's aus mit dem Trinken in den letzten Tagen?	-	-	-
34.	D2.1	Wie meinen Sie? (**A2.1:** Also haben Sie...) Akohol, oder?	v+	-	-
35.	A2.1	Nein, nein, normal Flüssigkeit.	-	-	IRNC
36.	D2.1	Vizavi de cum vă hidratați. Cât beți? În ultimele zile.	s+	-	-
37.	P2.1	Am băut normal, cam 1-1,5L, dar în weekend am fost într-o călătorie și nu am băut decat jumate de litru de apă toata ziua (**D2.1**: Aham).	-	-	IRNC
38.	D2.1	Ich hab 1 L bis eineinhalb Liter getrunken, aber am Wochenende war ich auf einem Ausflug und da habe ich den ganzen Tag nur einen halben Liter getrunken.	p	-	-
39.	A2.1	Und der Durchfall wie oft ist der?	-	-	DEMO
40.	D2.1	Diareea. Cât de des aveți diaree?	p	-	-
41.	P2.1	De 2-3 ori pe zi.	-	-	IRNC
42.	D2.1	2-3 Mal am Tag (**A2.1:** 2-3 Mal am Tag).	p	-	DISP of A2.1
43.	A2.1	Und auch nachts?	-	-	-
44.	D2.1	Și noaptea la fel?	p	-	-
45.	P2.1	Da, cât și în timpul nopții.	-	-	IRNC
46.	D2.1	Ja, während der Nacht auch.	p	-	-
47.	A2.1	Und wie ist die Farbe vom Stuhl?	-	-	-
48.	D2.1	Culoarea scaunului, care este?	p	-	-
49.	P2.1	Este gălbuie (**D2.1**: Aham).	-	-	ACK of D2.1
50.	D2.1	Gelblich.	p	-	-
51.	A2.1	Aham. Und haben Sie auch Schmerzen vor dem Stuhlgang oder während dem Stuhlgang?	-	-	ACK
52.	D2.1	Aveți dureri înainte de scaun sau în timpul defecației? În timpul eliberării scaunului?	s+	-	-
53.	P2.1	Înainte să am diaree mă doare foarte tare burta.	-	-	IRNC
54.	D2.1	Aham. Vor dem Stuhlgang habe ich ganz schlimme Bauchschmerzen.	p	-	ACK
55.	A2.1	Gut. Haben Sie in den letzten Tagen irgendwas gegessen, was außergewöhnlich war, zum Beispiel ein Hühnchen, Hendl von irgendwo, wo Sie es nicht selber gekocht haben?	-	-	ACK

56.	D2.1	Ați mâncat în ultima vreme ceva neobișnuit sau poate ați mâncat undeva carne de găina sau de pui pe care nu ați preparat-o dumneavoastră undeva în afara bucătăriei sau locurilor cunoscute?	p	-	-
57.	P2.1	Am mâncat în oraș inainte cu 2 zile.	-	-	IRNC
58.	D2.1	Vor 2 Tagen war ich in einem Lokal und hab dort gegessen. Also in der Stadt.	p	-	-
59.	A2.1	Aham. Ist jemand in der Umgebung krank?	-	-	ACK
60.	D2.1	S-a îmbolnăvit și altcineva din cercul dumneavoastră cunoscut?	p	-	-
61.	P2.1	Nu știu.	-	-	-
62.	D2.1	Davon weiß ich nix.	p	-	-
63.	A2.1	Okay. Gut, dann wegen dem Sprunggelenk, auf welcher Seite ist das?	-	-	ACK DEMO IRNC
64.	D2.1	Vizavi de glezne, [unde] pe ce parte se află?	p	-	-
65.	P2.1	Pe ce parte a gleznelor?	-	-	-
66.	D2.1	Sie möchten wissen, auf welche Seite der beiden Sprunggelenke?	v+	-	-
67.	A2.1	Genau, genau. Welches Sprunggelenk. Rechts oder links.	-	-	ACK
68.	D2.1	Exact, vreau să știu: glezna stângă sau glezna dreaptă?	p	-	
69.	P2.1	Ambele.	-	-	IRNC
70.	D2.1	Beide (**A2.1:** Beide).	p	-	DISP of A2.1
71.	A2.1	Okay. Hat ein Unfall stattgefunden? Ein Sturz oder ein Umknöcheln?	-	-	ACK
72.	D2.1	Ați avut cumva un accident? Poate v-ați sclintit piciorul sau ați căzut?	s	lc.g minor	-
73.	P2.1	Nu, doar acuma în weekend când am fost în seara accea m-au durut foarte tare gleznele după ce am umblat toata ziua.	-	-	IRNC
74.	D2.1	Nein, nur am Wochenende, wie gesagt, nach diesem Ausflug hatte ich schlimme Schmerzen der Spurengelenke, weil ich den ganzen Tag unterwegs war. Die Schmerzen waren dann am Abend.	s+	-	-
75.	A2.1	Okay. Sind die Sprunggelenke abgeschwollen?	-	-	ACK
76.	D2.1	Gleznele sunt umflate?	s	a.mt minor	-

77.	P2.1	Da.	-	-	IRNC
78.	D2.1	Ja.	p	-	-
79.	A2.1	Okay. Das werden wir uns dann gleich noch anschauen in der Untersuchung. Ja?	-	-	ACK
80.	D2.1	Ne vom uita la ele. Le vom controla.	p	-	-
81.	A2.1	Gibt's eine Allergie? Ist da irgendwas bekannt?	-	-	-
82.	D2.1	Aveți vreo alergie? Știți să aveți vreo alergie?	p	-	-
83.	P2.1	Din ce știu, nu am nicio alergie. N-am avut vreo reacție adversă la ceva.	-	-	IRNC
84.	D2.1	Keine bekannten Alergien. Bis jetzt hatte ich keine Nebenwirkungen.	P	-	-
85.	A2.1	Okay. Und in der Familie ist Richtung Darmerkrankungen oder Rheuma irgendwas bekannt?	-	-	ACK
86.	D2.1	[În familie] Vă este cunoscut să aveți pe cineva în familie care să aibă probleme gastrointestinale sau cu reumatism?	p	-	-
87.	P2.1	Doar bunica știu că nu poate să bea lapte.	-	-	IRNC
88.	D2.1	Ich weiß nur, dass meine Großmutter keine Milch trinken kann.	p	-	-
89.	A2.1	Bei Ihnen hat aber Milch bis jetzt nichts gemacht?	-	-	IRNC
90.	D2.1	Dar dumneavoastră nu aveți nimic de la lapte? (**A2.1:** Oder Milchprodukte?) Sau? (**P2.1:** Nu) Nu ați observat să aveți ceva de la produse lactate?	p	-	IRNC of P2.1
91.	P2.1	Nu am observat.	-	-	IRNC
92.	D2.1	Nein.	p	-	-

93.	A2.1	Gut, alles klar. So. Ich sage Ihnen jetzt kurz, was wir weitermachen. Ja? Also die Sachen gehören auf jeden Fall abgeklärt, ja? Und es [kommen jetzt gleich] kommt jetzt gleich die Pflege zu Ihnen, die wird Ihnen einen Zugang stechen, einen venösen Zugang und Blut abnehmen, ja? Damit wir schauen, ob zum Beispiel eine Entzündung im Körper irgendwo ist. Und es möglich ist, geben wir Ihnen so einen kleinen Becher, da bräuchten wir eine Stuhlprobe. Ja, dann können wir schauen, ob irgendwelche Keime, Bakterien, Viren drinnen sind, die diese Probleme verursachen könnten, ja? Es gibt bei einer Gastroenteritis, also bei diesem Magen-Darm-Virus, auch das, also ist oft so, dass zum Beispiel die Gelenke so reaktiv mitreagieren und, genau, dann müssen wir unbedingt ein Ultraschall machen, um uns die Harnblase anzuschauen, weil sie jetzt schon seit 2 Tagen keinen Harn mehr hatten. Ist das wirklich gar kein Harn?			ACK
94.	D2.1	O scurtă întrebare: după aceea, vă traduc ce a spus până acum. Vizavi de urinare, nu ați putut urina deloc în ultimele două zile? Absolut deloc?	v+	-	-
95.	P2.1	Am urinat în dimineața de după ce am revenit din călătorie și urina avea culoare destul de închisă și după aceea de 2 zile nu am mai urinat.	-	-	IRNC
96.	D2.1	Also, ich hab uriniert nach dem Ausflug am am nächsten Morgen, da war mein Urin sehr dunkelfarbig, danach konnte ich gar nicht mehr urinieren.	o	a.o minor	-
97.	A2.1	Okay, und im Unterbauch haben Sie Schmerzen, oder ist da nix?	-	-	ACK
98.	D2.1	Aveți în abdomenul inferior dureri, sau nu simțiți nimic?	p	-	-
99.	P2.1	Nu, doar atunci chiar înainte să am diaree.	-	-	IRNC
100.	D2.1	Nur kurz vor dem Durchfall. Ansonsten nichts.	p	-	-
101.	A2.1	Okay.	-	-	ACK
102.	D2.1	Ich muss ihm noch erklären was Sie noch planen.	v+	-	-
103.	A2.1	Ja klar.	-	-	ACK

104.	D2.1	Va veni imediat infirmiera și vă va pune o branulă, vă va lua sânge pentru a vedea dacă aveți vreo infecție în corp, după care vă vom da și un recipient mic. Avem nevoie de o probă de scaun pentru a vedea dacă sunt bacterii, viruși sau alte microorganisme care să fie cauza acestor probleme. Acești viruși gastrointestinali pot să ducă și la coreacția gleznelor și a încheieturilor. Vizavi de vezică, va trebui să facem un ecograf pentru a vedea ce se întâmplă acolo, având în vedere că de două zile nu ați putut urina.	s+	-	-
105.	P2.1	Okay. Mulțumesc, (**D2.1:** Ja, vielen Dank!) sper să nu fie ceva grav.	p	-	ACK of P2.1
106.	D2.1	Ich hoffe (**A2.1**: Aham) es wird nichts Schlimmes sein.	p	-	ACK of A2.1
107.	A2.1	Genau. Und ich werd Sie auch noch untersuchen, körperlich, und dann kriegen Sie ein bisschen Flüssigkeit über die Vene und Schmerzmittel und was gegen die Übelkeit. Ja?			
108.	D2.1	Vă voi controla fizic, după care veți primi și o substanță de hidratare intravenos și un analgezic împotriva stării de greață.	s-	a.mt major	-
109.	P2.1	Mulțumesc.	-	-	-
110.	D2.1	Vielen Dank (**A2.1:** Haben Sie noch Fragen akut?)!	p	-	-
111.	D2.1	Mei aveți întrebări acum acute?	p	-	-
112.	P2.1	Alte medicamente nu ar trebui să iau?	-	-	-
113.	D2.1	Müsste ich nicht auch andere Medikamente einnehmen?	p	-	-
114.	A2.1	Jetzt kriegen Sie mal die Akutmedikation und wenn'ma wissen, was Sie haben, dann schaun'ma wie die Therapie weitergeht.	-	-	IRNC
115.	D2.1	Întâi vă dăm medicamentele care se impun din cauza urgenței, după care, după ce elucidăm cauzele, vom vedea care este terapia de care aveți nevoie.	p	-	-
116.	P2.1	Okay.	-	-	ACK
117.	A2.1	Okay.	-	-	ACK

C4

Segment nr.	Speaker	Utterance	Analysis		
			Hamai et al. (2024)	MQM	Clark & Schaefer (1989)
1.	A2.2	Transcript Guten Tag, Dr. Heindlich ist mein Name. Was führt Sie zu uns in die Notaufnahme? Voice Guten Tag, Dr. Heindrich ist mein Name. Was führt Sie zu uns in die Notaufnahme?	-	-	-
2.	ChatGPT	Bună ziua, doctor Hendi îmi este numele. Ce vă aduce la noi în camera de urgență?	s	a.mt minor	-
3.	P2.2	Transcript Bună ziua, numel meu este Mihai Bacalbașa și am o durere foarte puternică în abdomen care a început de astăzi dimineață. Este aproape insuportabilă. Voice Bună ziua, numele meu este Mihai Bacalbașa și am o durere foarte puternică în abdomen care a început de astăzi dimineață. Este aproape insuportabilă.	-	-	IRNC
4.	ChatGPT	Guten Tag, mein Name ist Mihai Băcalbașa, und ich habe sehr starke Schmerzen im Bauch, die heute Morgen angefangen haben. Sie sind fast unerträglich.	s	a.mt minor	-
5.	A2.2	Auf einer Skala von 1 bis 10, wie sind die Schmerzen, also 1 ist kaum und 10 ist nicht zum Aushalten?	-	-	-
6.	ChatGPT	Pe o scară de la 1 la 10, cât de puternice sunt durerile, adică 1 înseamnă aproape deloc, iar 10 înseamnă insuportabile?	p	-	-
7.	P2.2	Transcript A zice că sunt nouă, a început de dimineață cu un jung, dar a început să mă doară tot mai tare pe parcursul zilei și am o apăsare foarte puternică și a ajuns să mă doare chiar până să simt și în spate. Voice Aș zice că sunt nouă, a început de	-	-	IRNC

		dimineață cu un junghi, dar a început să mă doară tot mai tare pe parcursul zilei și am o apăsare foarte puternică și o ajuns să mă doare chiar până să simt și în spate.			
8.	ChatGPT	Er sagt, dass es eine 9 ist, es hat heute Morgen mit einem Stechen begonnen, aber im Laufe des Tages wurde der Schmerz immer stärker, und es gibt einen sehr starken Druck, und der Schmerz strahlt sogar bis in den Rücken aus.	s	-	-
9.	A2.2	Und wo im Bauch ist der Schmerz, eher im Ober- oder im Unterbauch?	-	-	IRNC
10.	ChatGPT	Și unde se află durerea în abdomen, mai degrabă în partea de sus sau în partea de jos a abdomenului?	p	-	-
11.	P2.2	Transcript Este în partea de jos, ci încep s-o simt chiar aici, în spate. Nu știu cum, mă doare foarte tare. Voice Este în partea de jos și încep s-o simt chiar aicea, în spate. Nu știu cum. Mă doare foarte tare.	-	-	IRNC
12.	ChatGPT	Es ist im unteren Bereich, und ich beginne, es sogar hier im Rücken zu spüren, ich weiß nicht wie, es tut sehr weh.	p	-	-
13.	A2.2	Haben Sie schon ein Schmerzmittel eingenommen oder generell Medikamente?	-	-	IRNC of A2.2
14.	ChatGPT	Ați luat deja un analgezic sau, în general, medicamente?	p	-	-
15.	P2.2	Transcript Nu am luat încă niciun medicament, chiar acum am venit cu trenul și am și vomitat de două ori pe tren și am venit direct la spital după când am ajuns. Voice Nu am luat încă niciun medicament, [am avut] chiar acuma am ajuns, am venit cu trenul și am și vomitat de două ori pe tren și am venit direct la spital după când am ajuns.	-	-	IRNC
16.	ChatGPT	Nein, ich habe noch kein Medikament eingenommen, ich bin gerade erst angekommen, ich bin mit dem Zug gekommen und habe im Zug auch zweimal erbrochen, und ich bin direkt	p	-	-

		ins Krankenhaus gekommen, sobald ich angekommen bin.			
17.	A2.2	Transcript Haben Sie Fieber auch gehabt, also Schüttelfrost? Voice Okay, haben Sie Fieber auch gehabt, also Schüttelfrost?	-	-	ACK
18.	ChatGPT	Bine, ați avut și febră, adică frisoane?	p	-	-
19.	P2.2	Transcript - Voice Nu, dar simt și niște dureri în piept și îmi este foarte greu să respir.	-	-	IRNC
20.	ChatGPT	Transcript - Voice Nein, aber ich verspüre auch einige Brustschmerzen und es fällt mir sehr schwer zu atmen.	p	-	-
21.	A2.2	Sind Vorerkrankungen bekannt?	-	-	-
22.	ChatGPT	Sunt cunoscute antecedente medicale?	p	-	-
23.	P2.2	Transcript Am fost doar diagnosticat cu hipertleglicemie severă, dar nu mi-a fost prescris niciun medicament. Voice Am fost doar diagnosticat cu hipertriglicemie severă, [și] dar nu mi-a fost prescris niciun medicament.	-	-	IRNC
24.	ChatGPT	Ich wurde nur mit schwerer Hypertriglyzeridämie diagnostiziert, aber mir wurde kein Medikament verschrieben.	p	-	-
25.	A2.2	Haben Sie eine Allergie?	-	-	-
26.	ChatGPT	Aveți vreo alergie?	p	-	-
27.	P2.2	Nu, nu am alergii.	-	-	IRNC
28.	ChatGPT	Transcript Nein, ich habe keine Allergien. Voice (**A2.2:** Okay) Nein, ich hab…	p	s.as minor	ACK of A2.2
29.	A2.2	Trinken Sie regelmäßig Alkohol?	-	-	-
30.	ChatGPT	Consumați alcool în mod regulat?	p	-	-
31.	P2.2	Transcript - Voice Consum cam, să zic, 2 beri pe săptămână, dar săptămâna trecută am băut mai mult, deoarece am avut mai multe motive de a sărbători și am băut	-	-	IRNC

		câteva seri la rând mai mult și am băut în exces.			
32.	ChatGPT	Transcript Ich konsumiere ca. 2 Bier pro Woche, aber letzte Woche habe ich mehr getrunken, weil ich mehr Motive für Feierabend hatte. Ich habe einige Serien mehr getrunken und habe in Exzess getrunken. Voice Ich trinke etwa, sagen wir, zwei Biere pro Woche, aber in der letzten Woche habe ich mehr getrunken, weil ich mehrere Gründe zum Feiern hatte, und ich habe einige Abende hintereinander mehr getrunken und übermäßig konsumiert.	p	-	-
33.	A2.2	Gibt es beim Stuhlgang Veränderungen?	-	-	-
34.	ChatGPT	Există modificări la nivelul scaunului?	p	-	-
35.	P2.2	Transcript De când au început durere nu am fost încă la baie, deci nu am văzut. Voice De când au început durerile, nu am fost încă la baie, deci nu am văzut.	-	-	IRNC
36.	ChatGPT	Seitdem die Schmerzen begonnen haben, war ich noch nicht auf der Toilette, also habe ich es nicht gesehen.	p	-	-
37.	A2.2	Transcript Okay, gut. Es klingt so ein bisschen nach einer Bauchspeicheldrüsenentzündung. Das kann einmal durch einen vermehrten Alkoholkonsum ausgelöst werden. Wir werden Ihnen jetzt Blut abnehmen und die entsprechenden Parameter uns anschauen und geben Ihnen einmal Flüssigkeit und wir werden auch einen Ultraschall machen. Da schicken wir Sie zu den Radiologen. Voice Okay, gut. Es klingt so ein bisschen nach einer Bauchspeicheldrüsenentzündung. Das kann einmal durch einen vermehrten Alkoholkonsum ausgelöst werden. Ja? Wir werden Ihnen jetzt Blut abnehmen und die entsprechenden Parameter uns anschauen und geben Ihnen einmal	-	-	ACK

		Flüssigkeit und wir werden auch einen Ultraschall machen. Da schicken wir Sie zu den Radiologen.			
38.	ChatGPT	În regulă, bine, sună puțin a unei pancreatite, care poate fi declanșată de un consum crescut de alcool, da. Acum vă vom recolta sânge și vom analiza parametrii corespunzători, și vă vom administra și lichide, iar apoi vom efectua și o ecografie, vă vom trimite la radiologi.	p	a.as minor	-
39.	P2.2	Mulțumesc, am înțeles, dar puteți să-mi dați ceva pentru a-mi atenua durerile, deoarece mă doare foarte tare.	-	-	ACK
40.	ChatGPT	Vielen Dank, ich habe verstanden, aber können Sie mir etwas geben, um die Schmerzen zu lindern, denn es tut wirklich sehr weh.	p	-	-
41.	A2.2	Sie bekommen jetzt gleich was gegen die Schmerzen über die Vene und auch was gegen die Übelkeit.	-	-	IRNC
42.	ChatGPT	Acum o să vă administrăm imediat ceva pentru durere pe cale venoasă și, de asemenea, ceva împotriva greței.	p	-	-
43.	P2.2	Mulțumesc! Sper să pot să ajung la conferință. Nu credeam că o să-mi fie atât de rău.	-	-	ACK
44.	ChatGPT	Vielen Dank, ich hoffe, dass ich rechtzeitig zur Konferenz komme, ich hätte nicht gedacht, dass es so schlimm werden würde.	p	-	-
45.	A2.2	Transcript Wenn es wirklich eine Bauchspeicheldrüsenentzündung ist, bleiben Sie wahrscheinlich ein paar Tage im Spital. Voice Wir werden jetzt schauen, was los ist. Wenn es wirklich eine Bauchspeicheldrüsenentzündung ist, bleiben Sie wahrscheinlich ein paar Tage im Spital.	-	-	IRNC
46.	ChatGPT	Să vedem acum ce se întâmplă, dacă este într-adevăr o pancreatită, probabil veți rămâne câteva zile în spital.	p	-	-
47.	P2.2	Transcript Ce păcat, dar trebuie să scap de dureria asta, chiar este insuportabil. Voice	-	-	IRNC

		Ce păcat, dar trebuie să scap de durerea asta, chiar este insuportabil.			
48.	ChatGPT	Schade, aber ich muss diese Schmerzen loswerden, sie sind wirklich unerträglich.	p	-	-
49.	A2.2	Wir geben unser Bestes.	-	-	IRNC
50.	ChatGPT	Vom face tot posibilul.	p	-	-
51.	P2.2	Transcript - Voice Mulțumesc, doamnă doctor!	-	-	ACK of P2.2
52.	ChatGPT	Vielen Dank, wirklich!	p	a.mt minor	-

Interview Transcripts

Interview Patients Romanian

I: Am să încep atunci cu prima întrebare: Am vrut să vă întreb dacă aveți vreo experiență anterioară cu interpreți sau cu funcția aceasta de interpretare a ChatGPT?
P1: Eu nu am avut niciodată până acum.
P2: Cu interpreți profesioniști, nu.
I: Deci e ceva nou cam pentru toată lumea. Aș vrea să vă întreb cum ați descrie așa în mare comunicarea în timpul consultațiilor acestora?
P1: Păi, o să zic eu primul. Cu interpretul a fost foarte bine. Adică, nu știu, s-a simțit mult mai natural, confortabil. Pot să spun că am încredere că o să traducă cum trebuie. Doar că cu ChatGPT, eu personal nu am încredere în ce traduce ChatGPT sau, în mare parte, în ce îmi spune ChatGPT. Pentru că el nu e construit să-ți dea răspunsuri corecte. E construit pur și simplu să îți răspundă. Nu știu ce să zic mai mult. Clar sunt de partea interpretului fizic decât ChatGPT.
P2: Pentru mine, a fost o experiență okay. Pot să zic că cu un interpret pot să simți că este discuția mai cursivă, adică îți captează atenția. Parcă cu ChatGPT ai aceste momente de așteptare și trebuie să aștepți mult, să încarce, să înțeleagă. Câteodată la mine, de exemplu, vorbea destul de încet, câteodată sacadat sau avea câteodată un ton al vocii care nu se potrivea. Se schimba de la o conversație la alta.
I: În ce măsură credeți că medicul a înțeles problemele și preocupările dumneavoastră în ambele scenarii?
P1: Pot să zic de pe o scară de la 1 la 10?
I: Da, sigur. Poți să spui de la 1 la 10 pentru fiecare dintre scenarii.
P1: Pentru primul scenariu, 10 din 10 a înțeles tot ce am vrut să-i comunic. Dar cu ChatGPT 6 din 10. În primul rând, a dat informații greșite atunci când a spus că am febră de 63 de grade. Sau poate nu a tradus tot ce am spus eu, a tradus doar anumite propoziții din ce am spus.
I: P2, tu ce zici? Traducerea cu interpret mi se pare că a fost foarte bună. S-a văzut că a fost într-un cadru profesionist și nu au fost probleme. La ChatGPT, poate din cauza scenariului meu sau din cum am comunicat, n-am observat probleme. Nici nu am înțeles ce a tradus, deci nu pot să-mi dau cu părerea dacă a înțeles medicul sau nu. Din ce mi-a transmis înapoi, am înțeles că a înțeles, adică mesajele principale le-a transmis și medicul mi-a transmis informația înapoi la

fiecare întrebare sau fiecare replică pe care am avut-o. Doar mi se părea că partea cu vorbitul sacadat sau tonul pe care îl foloseam mi se părea foarte ciudat.

I: Deci ați spune că ați înțeles întrebările, explicațiile, instrucțiunile medicului în ambele scenarii?

P1: Da.

P2: Da.

I: Au existat momente în care n-ați fost siguri ce a vrut să spună medicul în oricare dintre scenarii?

P2: Nu.

P1: Nu am înțeles ce a vrut să spună medical.

I: Momente de confuzie au existat? Fie ele tehnice sau de limbă?

P2: La scenariul meu, a fost doar când n-am înțeles partea cu gleznele, care gleznă, dar nu știu care a fost problema. Poate a fost cum am înțeles eu.

P1: Și pentru mine a fost în scenariul cu ChatGPT puțină confuzie.

I: Știi mai exact într-un moment anume sau în general?

P1: Atunci când, de exemplu, doctorul trebuia să spună ceva, a spus, ChatGPT n-a zis nimic și după ea a repetat, apoi ChatGPT a întrerupt-o, ea a început să vorbească, nu știu dacă a tradus ce a trebuit.

I: Considerați că unul dintre modurile acestea a fost mai util decât celălalt? Mai la îndemână? Mai favorabil?

P2: Pentru a fi siguri de ceea ce se întâmplă și pentru a avea mai multă încredere că pot să rezolv problema rapid, poate e mai bine cu un interpret, aș zice eu. Pentru că, de exemplu, cum aș vedea: te duci cu un telefon să ai o discuție cu medicul; dacă ai probleme foarte grave, până aștepți să-ți traducă...

P1: Da, sunt de acord și eu.

P2: Și poate din cauză că ai dureri, nu poți să vorbești. De exemplu, nu știu dacă ai dureri și vorbești foarte întrerupt sau foarte distorsionat din cauză că nu te poți concentra. Ar înțelege foarte bine.

I: Înseamnă că considerați că interpretul a transmis mai bine mesajul dumneavoastră. Ce credeți că a contribuit la acest lucru?

P1: Interpretul este om, poate să mă înțeleagă. Adică, bineînțeles, depinde și de aptitudinea interpretului. Dar cu un interpret poți să fii sigur că ceea ce ai tradus e transmis doctorului. Cu ChatGPT nu poți fi foarte sigur.

P2: De exemplu, nu sunt foarte sigur când m-am exprimat foarte... poate nu corect, sau am avut exprimări greșite, dacă prin ChatGPT s-a înțeles exact la ce mă refer sau a trebuit să deducă sau se poate să fie unele derivări. Dar cred că un interpret a putut să facă acest… adică să-mi comunice într-un mod mai corect ceea ce am vrut să spun.

I: Adică să înțeleagă din context?

P2: Da.

P1: De obicei, în ChatGPT o ia pe arătură.

I: Cum credeți că a afectat modul de interpretare fluxul conversațional și emoțional al consultației?

P1: Păi... emoțional, ce să zic, interpretul poate să comunice emoțiile mele mai bine către doctor, cred, decât ChatGPT.

P2: Dar nici eu n-am simțit că ChatGPT putea să transmită foarte bine faptul că, de exemplu, am o urgență, că mă doare sau ce probleme am neapărat în momentul acela. Și acele momente de pauză pe care trebuie să le ai ca să proceseze și să înțeleagă, să treacă de la o replică la alta; cu un interpret e mult mai rapid și, de exemplu, poți să te uiți în ochii lor și să vezi, să comunici. E mult mai natural.

P1: Da.

I: Dar ați menționat că vi s-a părut că nu a avut tonul potrivit, deci practic a transmis o emoție și ChatGPT.

P1: Da, dar nu emoția care trebuie.

I: Deci, ați simțit că nu e aceeași linie pe care ați mers și voi?

P2: Nu.

P1: Nu, nu. Cu mine a fost agresiv.

I: Ați schimbat ceva în modul în care vorbiți de obicei, ca să facilitați conversația fie cu interpretul, fie cu ChatGPT?

P1: Nu, nu cred. Eu nu.

P2: Eu zic că cu ChatGPT ar trebui să fii mult mai precis și să nu folosești fraze foarte lungi. Eu asta am încercat, ca să fie mai cursiv.

I: Deci, totuși, te-ai gândit cum să formulezi mesajul înainte să-l dai?

P2: Da, adică să fie mai scurt, m-am gândit. Până traducea, până se gândea, am zis că mai bine să fie un mesaj scurt.

I: V-ați simțit, ascultați sau înțeleși de către doctor prin ambele medii sau printr-unul mai mult, printr-altul mai puțin?

P1: Per total m-am simțit înțeles în ambele feluri, adică cu ambele moduri. Cred că am transmis mesajul și cred că tratamentul ar fi fost la fel sau diagnosticul cred că ar fi fost la fel în ambele cazuri.

P2: Asta poate să depindă și de medic și de priceperea așa a medicului de a folosi aceste tooluri, de a înțelege și de a lucra.

P1: Adică, da, depinde și de medic.

I: Cât de empatic vi s-a părut medicul în timpul fiecărei consultații? Vi s-a părut că ar fi diferit în cele două scenarii sau la fel de empatic?

P1: Mie mi s-a părut la fel în ambele scenarii.

P2: Da, la fel. În cazul medicului n-am observat diferențe.

I: Și vi se pare important acest lucru? Nivelul acesta de empatie? Sau nu e atât de important? Mai importantă e informația?

P2: Da, ambele sunt importante. Empatie, mi se pare, poți să transmiți față de medic mai ușor prin interpret, dar când vine vorba de a simți înapoi, depinde doar de medic, nu depinde neapărat de interpret.

P1: Da, eu cred că medicul ar trebui să se comporte la fel în ambele situații.

I: A existat vreo reacție neașteptată din partea medicului care să vă indice că n-a înțeles sau o discrepanță între ceea ce ați spus și ceea ce s-a înțeles prin interpretare?

P1: Da, la un moment dat pentru mine. În acel moment când ChatGPT a tradus și a durat foarte mult timp să vorbească și doctorul a început să vorbească între timp, ChatGPT l-a întrerupt, în fine, sau când ChatGPT pur și simplu a dat informații greșite, febra, atunci a fost un moment vizibil de confuzie.

P2: La mine nu am avut.

I: Deci mi-ați spus deja că ați observat că ChatGPT ar fi omis, de exemplu, anumite informații pentru că a întrerupt medicul. Deci, vi s-a părut că ar fi situații de genul acesta mai des sau doar aceste întreruperi? Au existat omisiuni, modificări? Ați simțit lucrul acesta?

P1: Per total, nu foarte multe care să afecteze diagnosticul final, deci cred că e okay.

P2: Da, comparativ, ca impresie, am impresia că ChatGPT ar fi avut mai multe greșeli și ar fi fost mai greu de înțeles, doar așa, per total, dar nu a fost o diferență mare.

I: Deja am vorbit despre următoarea întrebare, despre ton, dar vreau să vă mai întreb dacă ați avut impresia că gradul de severitate sau conținutul emoțional a fost transmis în ambele cazuri?

P2: În cazul meu, mi se pare că a fost transmis greșit gradul de severitate. Așa mi s-a părut mie că sună când a tradus ceea ce vreau eu să spun.

I: În ce caz?

P2: În cazul când am comunicat cu ChatGPT.

P1:Pentru mine, cred că interpretul a transmis corect, dar ChatGPT nu cred că a transmis gradul de severitate.

I: Dar ce te face să ai această impresie?

P1: Doar cum a vorbit ChatGPT, cum s-a exprimat.

I: Ați sesizat vreo diferență în modul în care a fost utilizată terminologia medicală în cele două cazuri?

P1: Nu, eu nu am remarcat.

P2: Nu, n-am remarcat nimic.

I: Cum au afectat erorile acestea de care îmi spuneți la ChatGPT fluxul conversației? Ce s-a întâmplat? A fost nevoie să repetați? Cum a schimbat fluxul conversației?

P1: Eu am repetat niște informații din cauza greșelilor. Și când am văzut clar că doctorul e confuz, că nu se așteptă răspunsurile lui ChatGPT.

P2: La mine doar mi s-a părut conversația mai greoaie din cauza momentelor de așteptare și faptului că, în acel moment, când vorbea destul de sacadat, trebuia să te concentrezi pe ceea ce spunea, ca să poți să înțelegi. În rest, cred că vorbea corect și am înțeles termenii, dar asta a afectat un pic conversația.

I: Mai există ceva ce ați observat și credeți că ar fi important să menționați?

P1: Nu, în cazul meu nu.

P2: Nu, nu. Cred că am exprimat cam toată experiența.

I: Nu v-a deranjat nimic în cele două scenarii în afară de ceea ce mi-ați spus deja?

P1: Nu.

P2: Nu.

I: Bine, atunci vă mulțumesc mult!

Interview Patients English

I: I will start with the first question: I wanted to ask if you have any previous experience with interpreters or with this ChatGPT interpreting function?

P1: I have never had any experience with interpreters or with the ChatGPT interpreting function before.

P2: With professional interpreters, no.

I: So it is something new for everyone. I would like to ask you how you would describe the communication during these consultations in general?

P1: Well, I will go first. It was very good with the interpreter. I mean, I do not know, it felt much more natural, comfortable. I can say that I trust that they will translate correctly. But with ChatGPT, I personally do not trust what ChatGPT translates or, for the most part, what ChatGPT tells me. Because it is not designed to give you correct answers. It is simply designed to respond to you. I do not know what else to say. I am definitely on the side of the physical interpreter rather than ChatGPT.

P2: For me, it was an okay experience. I can say that with an interpreter, you can feel that the conversation is more fluid, meaning it captures your attention. With ChatGPT, it is like you have these moments of waiting, and you have to wait a long time for it to load and understand. Sometimes, for example, it spoke quite slowly, sometimes haltingly, or sometimes had a tone of voice that did not fit. It changed from one conversation to another.

I: To what extent do you think the doctor understood your problems and concerns in both scenarios?

P1: Can I say on a scale of 1 to 10?

I: Yes, of course. You can say from 1 to 10 for each of the scenarios.

P1: For the first scenario, 10 out of 10, it understood everything I wanted to communicate. But with ChatGPT, 6 out of 10. First of all, it gave wrong information when it said I had a fever of 63 degrees. Or maybe it did not translate everything I said; it only translated certain sentences from what I said.

I: P2, what do you think?

P2: I think the translation with the interpreter was very good. It was clear that it was in a professional setting, and there were no problems. With ChatGPT, maybe because of my scenario or how I communicated, I did not notice any problems. I also did not understand what it translated, so I cannot say whether the doctor understood or not. From what they said to me, I gathered that they understood, meaning it conveyed the main message, and the doctor

answered each of my questions or remarks. It just seemed to me that the stuttered speech or the tone it was using seemed very strange.

I: So would you say that you understood the doctor's questions, explanations and instructions in both scenarios?

P1: Yes.

P2: Yes.

I: Were there any moments when you were not sure what the doctor meant in any of the scenarios?

P2: No.

P1: No, I did understand what the doctor meant.

I: Were there moments of confusion, whether technical or linguistic?

P2: In my scenario, it was only when I did not understand the part about the ankles, which ankle, but I do not know what the problem was. Maybe it was because I did not understand what was meant.

P1: I also found the ChatGPT scenario a little confusing.

I: Do you know exactly when, or was it just in general?

P1: When, for example, the doctor had to say something, they said it; ChatGPT did not say anything, and then they repeated it. Then ChatGPT interrupted her. They started talking. I do not know if it translated what it was supposed to.

I: Do you think one of these methods was more useful than the other? More convenient? More favourable?

P2: To be sure of what is happening and to have more confidence that I can solve the problem quickly, maybe it is better with an interpreter, I would say. Because, for example, how I would see it: you go with a phone to have a discussion with the doctor; if you have very serious problems, you then have to also wait for the translation...

P1: Yes, I agree.

P2: And maybe because you are in pain, you cannot talk. For example, I do not know, maybe if you are in pain and you are talking very haltingly or very garbled because you cannot concentrate. [I am not sure if] It would understand very well.

I: So I gather you think the interpreter conveyed your message better. What do you think contributed to this?

P1: The interpreter is human, and they can understand me. I mean, of course, it also depends on the interpreter's skills. But with an interpreter, you can be sure that what you have translated is conveyed to the doctor. With ChatGPT, you cannot be very sure.

P2: For example, I am not very sure when I expressed myself very... perhaps not correctly, or I used the wrong words, whether I have made myself clear through ChatGPT, and it was understood exactly what I meant, or one had to deduce it, or there may have been some misinterpretations. But I think the interpreter was able to do this... that is, to communicate what I wanted to say more accurately.

I: Meaning to understand from the context?

P2: Yes.

P1: Usually, ChatGPT goes off the rails.

I: How do you think the conversational and emotional flow of the consultation affected the interpretation?

P1: Well... emotionally, what can I say? I think the interpreter can communicate my emotions to the doctor better than ChatGPT.

P2: But I did not feel that ChatGPT could convey very well that, for example, I had an emergency, that I was in pain or what problems I was having at that moment. And those pauses you have to take for it to process and understand, to move from one reply to another; with an interpreter, it is much faster and, for example, you can look them in the eye and see, communicate. It is much more natural.

P1: Yes.

I: But you mentioned that you felt it did not have the right tone, so ChatGPT also conveyed an emotion.

P1: Yes, but not the right emotion.

I: So you felt that it was not on the same wavelength?

P2: No.

P1: No, no. It was aggressive with me.

I: Did you change anything in the way you usually speak to facilitate the conversation with either the interpreter or ChatGPT?

P1: No, I do not think so. I did not.

P2: I think with ChatGPT, you should be much more precise and not use very long sentences. That is what I tried to make it more fluent.

I: So, you still thought about how to phrase the message before sending it?

P2: Yes, I thought it should be shorter. It took time to translate, to think, and I thought it would be better to have a short message.

I: Did you feel listened to or understood by the doctor through both channels or through one more than the other?

P1: Overall, I felt understood in both ways, that is, with both methods. I think I got the message across, and I think the treatment would have been the same or the diagnosis would have been the same in both cases.

P2: That may also depend on the doctor and the doctor's ability to use these tools, to understand and to work.

P1: Yes, it also depends on the doctor.

I: How empathetic did you find the doctor during each consultation? Did you find them to be different in the two scenarios or equally empathetic?

P1: I found them to be the same in both scenarios.

P2: Yes, the same. I did not notice any differences in the doctor's case.

I: And do you think this is important? This level of empathy? Or is it not that important? Is the information more important?

P2: Yes, both are important. Empathy, I think, can be conveyed to the doctor more easily through an interpreter, but when it comes to feeling it back, it depends only on the doctor, not necessarily on the interpreter.

P1: Yes, I think the doctor should behave the same in both situations.

I: Was there any unexpected reaction from the doctor that indicated that they did not understand, or a discrepancy between what you said and what was understood through interpretation?

P1: Yes, at one point for me. At the moment when ChatGPT translated, and it took a long time to start speaking, and the doctor started talking in the meantime, then ChatGPT interrupted them, anyway. Or when ChatGPT simply gave wrong information, regarding the fever, then there was a visible moment of confusion.

P2: For me, it was not the case.

I: So you already told me that you noticed that ChatGPT would have omitted, for example, certain information because it interrupted the doctor. So, did you feel that there were situations like this more often or just these interruptions? Were there omissions, changes? Did you get this impression?

P1: Overall, not many that would affect the final diagnosis, so I think it is okay.

P2: Yes, comparatively, as an impression, I feel that ChatGPT would have had more mistakes and would have been more difficult to understand overall, but it was not a big difference.

I: We have already talked about the next question, about tone, but I want to ask you if you felt that the degree of severity or emotional content was conveyed in both cases?

P2: In my case, I think the severity was conveyed incorrectly. That is how it sounded to me when it translated what I wanted to say.

I: In which case?

P2: When I communicated with ChatGPT.

P1: For me, I think the interpreter conveyed it correctly, but I do not think ChatGPT conveyed the severity.

I: But what makes you think that?

P1: Just the way ChatGPT spoke, the way it expressed itself.

I: Did you notice any difference in the way medical terminology was used in the two cases?

P1: No, I did not notice any.

P2: No, I did not notice anything.

I: How did these errors you mentioned from ChatGPT affect the flow of the conversation? What happened? Did you have to repeat yourself? How did it change the flow of the conversation?

P1: I repeated some information because of the mistakes, and when I saw clearly that the doctor was confused, that they were not expecting ChatGPT's answers.

P2: For me, the conversation just seemed a bit awkward because of the pauses and the fact that, at that moment, when it was speaking quite haltingly, you had to concentrate on what it was saying in order to understand. Otherwise, I think it spoke correctly, and I understood the terms, but that affected the conversation a little.

I: Is there anything else you noticed that you think is important to mention?

P1: No, not in my case.

P2: No, no. I think I have covered pretty much the whole experience.

I: Was there anything else that bothered you in the two scenarios apart from what you have already told me?

P1: No.

P2: No.

I: Okay, thank you very much!

Interview Healthcare Provider German

I: Ich möchte erst mal fragen, ob du schon mal Erfahrung mit Dolmetscher:innen oder mit der Dolmetschfunktion von ChatGPT hattest?
A: Ich glaube so einen richtig professionellen Dolmetscher oder Dolmetscherin, hatte ich noch nie, aber heute zum Beispiel war ein Sozialarbeiter mit einem polnischen Patienten mit, der übersetzt hat. Aber das war jetzt im Vergleich zu der Dolmetscherin nur mündlich, also er hat sich nicht irgendwie was mitgeschrieben, eher so, und natürlich viele Angehörige, die übersetzen, für die Patienten. Und ChatGPT habe ich noch nie benutzt. Eher Google, habe was was angegeben und die Patienten haben gelesen oder das auf laut geschaltet.
I: Also eher eingegeben und schriftlich und zum Lesen gegeben oder zum Vorlesen?
A: Genau, vor allem schriftlich, ja. Ich weiß, es gibt von Google Gemini, was eine Patientin zuletzt benutzt hat, da hat sie mir das hingehalten, ich habe was gesagt und das hat es ihr, ich glaube, nur schriftlich übersetzt.
I: Okay, also du hattest schon irgendwie Erfahrungen entweder mit. Dolmetscher:innen, Angehörige oder mit Technologie. Und so allgemein, wie ist dein Eindruck von diesen Konsultationen? Wie würdest du sie jetzt so kurz beschreiben?
A: Meinst du jetzt von den Szenarien von heute?
I: Ja genau, von den Szenarien.
A: Ich fand, es hat eigentlich ganz gut funktioniert. Ich hoffe, ich habe überall herausgefunden, was die Patienten haben. Angenehmer war es mit der Dolmetscherin, weil da ging es eben flüssiger, wie ein Gespräch und mit ChatGPT muss man immer warten; Übersetzt es jetzt? Und ich weiß auch nicht, ob es richtig übersetzt, was ich eigentlich gesagt habe. Ja, und eine Person ist halt immer zuverlässiger als der Computer.
I: Okay, wie gut denkst du, dass du die gesundheitlichen Beschwerden und Bedenken der Patient:innen verstanden hast in den beiden Szenarien?
A: Ich glaube, eigentlich bei beiden Szenarien ganz gut. Es hat eigentlich gepasst.
I: Okay, und hattest du auch das Gefühl, dass die Patient:innen deine Fragen und Erklärungen verstehen und nachvollziehen können?
A: Hatte ich eigentlich schon das Gefühl, ja, auch bei beiden. Bei beiden Szenarien.
I: Gab es Momente, in denen du nicht sicher warst, was die Patient:innen meinten?
A: Ja, einmal bei der Temperatur, da war ich mir nicht sicher, ob es in Rumänien vielleicht Fahrenheit gibt. Ich glaube, das ist nur in Amerika. Und da habe ich auch versucht mal nachzufragen, aber da war, glaube ich, einfach ein Fehler. Also das hat ChatGPT falsch

irgendwie übersetzt, aber ansonsten hatte ich auch immer das Gefühl, dass wir uns verstanden haben.

I: Also, es gab schon Teile, die verwirrend waren in den Konsultationen oder war das dann gelöst?

A: Ja, ja, aber ich glaube, das wurde ganz gut gelöst.

I: Okay. Genau, glaubst du, dass einer der Modi hilfreicher war als der andere, wenn ja, welcher?

A: Also, ich persönlich würde die Dolmetscherin einfach präferieren. Weil auch ebenso Nachfragen zu Kleinigkeiten sicher besser geht, aber das Handy und ChatGPT hat man halt schneller. Ich glaube, man kann beide Sachen gut verwenden. Also besser als dass der Patient überhaupt nicht versteht, was ich will, und ich nicht verstehe, was der Patient hat.

I: Und welcher Dolmetschmodus hat die Botschaft am besten vermittelt? Was meinst du?

A: Ja, ich glaube die Dolmetscherin.

I: Was hat dazu beigetragen, dass die Dolmetscherin das besser gemacht hat?

A: Sie kann halt auch direkt nachfragen, wenn sie was nicht verstanden hat, oder wenn sie was von mir nicht versteht, wenn sie was vom Patienten nicht versteht. Das geht einfach besser und auch so eine irgendwie das Zwischenmenschliche kommt doch besser rüber.

I: Und wie hat die Art der Verdolmetschung den Gesprächsfluss und die emotionale Atmosphäre der Konsultationen beeinflusst?

A: Schwierige Frage.

I: Vielleicht konzentrieren wir uns erstmal auf einzelne Aspekte, vielleicht erst mal auf den Gesprächsfluss.

A: Ähm, das war gut, also es war gut, weil sie hat das gesagt, was ich sagen wollte. Also so hat es gewirkt, als würde sie komplett alles aufnehmen und das hat eigentlich gut funktioniert. Und bei ChatGPT ist das halt immer, ich habe dann immer so das Gefühl, Okay, darfst du nicht zu viel sagen, weil sonst nimmt es das vielleicht nicht, und das es eben alles so wiedergibt. Weil sie kann mich halt auch unterbrechen, wenn sie sagt: Okay, jetzt würde sie es einmal weitergeben. Und bei ChatGPT, wenn man dann eine Pause dazwischen macht, dann hört es vielleicht auf zum Aufnehmen und da bin ich ein bisschen unsicher mit dem.

I: Und die emotionale Atmosphäre, also der emotionale Charakter des Gesprächs, wie war dein Eindruck da?

A: Es ist irgendwie leichter, wenn man jemanden, also wenn das wirklich alles menschlich ist. Klar, da ist noch eine Person und dann ist man irgendwie auch nervös, weil man auch nicht weiß, dass die von einem denkt und so. Das ist vielleicht meine eigene Meinung. Und vielleicht

auch im Szenario geschuldet. Aber ja, man ist sich irgendwie sicherer dann einfach, dass das alles klappt als mit dem Handy. Da ist man halt mit dem Patienten alleine. Also ich finde es besser mit dem Dolmetscher.

I: Also man hat dann eine Unterstützung quasi?

A: Ja, genau. Vielleicht können dann auch die Emotionen besser rübergebracht werden.

I: Hast du was an deiner üblichen Sprechweise geändert in den Szenarien?

A: Ich glaube, bei ChatGPT habe ich versucht, noch deutlicher zu sprechen und langsamer. Ich weiß nicht, warum es mal nicht verstanden hat und … Ich meine, ich habe keinen Dialekt, aber halt eher noch 'hochdeutscher', würde ich sagen, als jetzt mit der Dolmetscherin.

I: Okay, hat eines der Dolmetschmodi deine Fähigkeit, eine Verbindung zum Patienten oder zur Patientin aufzubauen, verbessert oder beeinträchtigt?

A: Ich glaube, es war bei beiden in Ordnung.

I: Okay, also das Übertragungsmedium spielt jetzt keine Rolle.

A: Nicht wirklich, nein.

I: Und glaubst du, dass die Patient:innen deine Empathie durch die Dolmetscherin als auch, durch das S2S System wahrgenommen haben?

A: Meine Sprachmelodie und so haben Sie ja alle bemerkt, aber die Stimme von ChatGPT ist halt sehr empathielos. Das ist zumindest im Deutschen, ich weiß nicht, wie es im Rumänischen ist. Das kommt sicher besser durch die Dolmetscherin rüber. Aber so, wie es von den Antworten her war, glaube ich, ist es bei beiden Szenarien gut rübergekommen; zumindest hoffe ich, dass ich empatievoll war.

I: Also, das ist dir auch wichtig? Dass du so wirkst.

A: Ja, ja, schon.

I: Gab es unerwartete Reaktionen von den Patient:innen, die auf eine Diskrepanz zwischen dem, was du gesagt hast, und dem, was sie durch die Dolmetschung verstanden haben, hindeuten könnten?

A: Hatte ich eigentlich nicht das Gefühl. Mir ist so vorgekommen, als würde das Ganze so ganz gut hin und her gehen.

I: Okay, hast du bemerkt, dass irgendetwas ausgelassen wurde, geändert wurde, hinzugefügt wurde?

A: Ich hätte es nicht bemerkt. Ich glaube, ich hoffe, es ist noch nicht passiert, also bei der Dolmetscherin hatte ich wirklich das Gefühl, sie hat alles. Also sie 2 Seiten teilweise mitgeschrieben und das dann auch wirklich lange übersetzt. Bei ChatGPT kann ich nicht sagen, aber eigentlich, so von den Antworten hat es immer ganz gut zusammengepasst.

I: Und hattest du auch das Gefühl, dass der Ton oder die Dringlichkeit, der emotionale Inhalt der Botschaft korrekt vermittelt wurde, oder hattest du das Gefühl, dass es irgendwie anders rübergekommen ist?

A: Nein, ich glaube, das ist auch so rübergekommen, wie es sein soll.

I: Hast du irgendwelche Unterschiede in der Verwendung von medizinischer Fachterminologie gemerkt? Zwischen den zwei Modi.

A: Also beim ChatGPT, das ist mir mehr aufgefallen. Ja, ich meine beim Rumänischen, da höre ich halt nur die medizinischen Sachen raus, weil die halt sehr, wie das Lateinische, klingen. Aber nein, eigentlich nicht.

I: Okay, und falls du Genauigkeitsfehler oder irgendwelche Fehler bemerkt hast, wie hat sich das auf den Ablauf der Konsultation ausgewirkt?

A: Habe ich jetzt eigentlich auch nicht so bemerkt, ich habe halt dann versucht noch mal nachzufragen, wo irgendwas unklar war, das sind dann immer die Antworten rübergekommen, wie ich sie mir erhofft habe. Eigentlich auch bei beiden.

I: Okay. Das hast du mir schon beantwortet, dass du die Dolmetscherin bevorzugen würdest, wenn du das noch mal machen würdest. Aber gibt es vielleicht noch etwas, was dir aufgefallen ist, was du noch erwähnen wollen würdest?

A: Ich finde es immer ein bisschen schwierig beim Dolmetschen, wie man dann spricht. Spricht man dann mit der Dolmetscherin über den Patienten oder spricht man immer den Patienten direkt an? Und ich weiß ja auch nicht, wie sie es übersetzt, also sie wird ihn direkt ansprechen, aber für euch, wie ist das besser? Ich frag sie über ihn oder sie oder ich spreche einfach direkt, als würde ich wirklich mit den Patienten sprechen, und sie ist wirklich nur das Medium. Also da bin ich ein bisschen unsicher. Weil bei ChatGPT ich halt wirklich fragen kann: Wie geht es Ihnen? Bei Dolmetscher:innen weiß ich nicht immer, wie ich das machen soll. Ich glaube, ich bin auch sehr oft geswitcht.

I: Ja, das ist eigentlich sehr interessant. Ich kann dir vielleicht verraten, dass wir es an der Uni so als quasi ‚Übertragungsmedium' machen und wir die Patient:innen anreden, als ob wir die Ärzt:innen sind, und dich auch, als ob wir die Patient:innen wären. Aber das machen auch nicht alle Dolmetscher:innen. Manche reden einfach in der dritten Person. Und zum Beispiel aus meiner Erfahrung reden die meisten Ärzt:innen. Mit mir, mit den Patient:innen.

A: Ja, da liegt auch die Gefahr. Weil ich mich ja nur mit ihr quasi unterhalten und auch eine Antwort bekommen kann, ist es ja eigentlich schon wichtig, dass der Patient auch sieht, okay, ich rede eigentlich mit ihm. Das ist schwierig, aber hm ich hoffe das war nicht so schlimm.

I: Das war super. Danke dir!

Interview Healthcare Provider English

I: First of all, I would like to ask if you have ever had any experience with interpreters or with ChatGPT's interpreting function?

A: I do not think I have ever had a really professional interpreter, but today, for example, there was a social worker with a Polish patient who translated. But compared to the interpreter, it was only verbal, so they did not write anything down, and of course, many relatives who translate for the patients. And I have never used ChatGPT. I have rather used Google, typed something in, and the patients read it or played it aloud.

I: So you type it in and give it to them to read or to play it aloud?

A: Exactly, mainly in writing, yes. I know there is Google Gemini, which a patient used recently. They held it up to me, I said something, and it translated it for them, I think only in writing.

I: Okay, so you have had some experience with either interpreters, relatives or technology. And in general, what is your impression of these consultations? How would you describe them briefly?

A: Do you mean today's scenarios?

I: Yes, exactly, the scenarios.

A: I thought it worked quite well, actually. I hope I figured out what the patients had in all situations. It was more pleasant with the interpreter because it was smoother, like a conversation, and with ChatGPT, you always have to wait; is it translating now? And I do not know if it is translating what I actually said correctly. Yes, and a person is always more reliable than a computer.

I: Okay, how well do you think you understood the patients' health complaints and concerns in the two scenarios?

A: I think quite well in both scenarios, actually. It worked out fine.

I: Okay, and did you also feel that the patients understood and could follow your questions and explanations?

A: I did have that feeling, yes, with both of them. In both scenarios.

I: Were there moments when you were not sure what the patients meant?

A: Yes, once with the temperature, I was not sure if they use Fahrenheit in Romania. I think that is only in America. And I tried to ask about it, but I think it was just a mistake. So ChatGPT translated it wrong somehow, but otherwise, I always felt that we understood each other.

I: So, were there parts of the consultations that were confusing, or was that all resolved?

A: Yes, yes, but I think it was resolved quite well.

I: Okay. Right, do you think one of the methods was more helpful than the other, and if so, which one?

A: Well, personally, I would prefer the interpreter. Because it is easier to ask questions about small details, but the mobile phone and ChatGPT are more easily accessible. I think both can be used effectively. It is better than the patient not understanding what I want at all, and me not understanding what the patient is suffering from.

I: And which interpreting method conveyed the message best? What do you think?

A: Yes, I think the interpreter.

I: What contributed to the interpreter doing a better job?

A: They can ask questions directly if they do not understand something, or if they do not understand something I say, or if they do not understand something the patient says. It just works better, and somehow the interpersonal aspect comes across better.

I: And how did the method of interpreting influence the flow of conversation and the emotional atmosphere of the consultations?

A: That is a difficult question.

I: Maybe we should focus on individual aspects first, perhaps the flow of conversation.

A: Um, that was good, I mean, it was good because they [the interpreter] said what I wanted to say. So it seemed as if they were taking in everything, and that actually worked well. With ChatGPT, I always have this feeling that I cannot say too much, because otherwise it might not pick it up and reproduce everything exactly. Because the interpreter can also interrupt me when they say: Okay, now they would like to translate. And with ChatGPT, if you pause in between, it might stop recording, and I am a bit unsure about that.

I: And the emotional atmosphere, that is, the emotional character of the conversation, what was your impression there?

A: It is somehow easier when you are talking to someone, when it is all human. Of course, there is another person there, and then you are somehow nervous because you do not know what they think of you and so on. That is perhaps my own opinion. And perhaps also due to the scenario. But yes, you are somehow more confident that everything will work out with the interpreter than with the mobile phone. With the mobile phone, you are alone with the patient. So I think it is better with the interpreter.

I: So you have support, so to speak?

A: Yes, exactly. Maybe it is easier to convey emotions that way.

I: Did you change anything about the way you usually speak in the scenarios?

A: I think with ChatGPT I tried to speak even more clearly and more slowly. I do not know why it did not understand sometimes, and ... I mean, I do not have a dialect, but I would say I spoke more 'standard German' than I did with the interpreter.

I: Okay, did either of the interpreting methods improve or impair your ability to connect with the patient?

A: I think it was fine with both.

I: Okay, so the transmission medium does not matter now.

A: Not really, no.

I: And do you think the patients perceived your empathy through the interpreter as well as through the S2S system?

A: They all noticed my intonation and so on, but ChatGPT's voice is very unempathetic. At least in German, I do not know how it is in Romanian. That certainly comes across better through the interpreter. But judging by the responses, I think it came across well in both scenarios; at least I hope I was empathetic.

I: So, is that important to you too? That you come across that way.

A: Yes, yes, it is.

I: Were there any unexpected reactions from patients that might indicate a discrepancy between what you said and what they understood through the interpretation?

A: I did not really get that impression. It seemed to me that everything was going quite well.

I: Okay, did you notice anything being left out, changed or added?

A: I did not notice. I think, I hope it also did not happen, because I really felt that the interpreter had everything. They wrote down two pages and then translated them, which took a really long time. I cannot say about ChatGPT, but actually, from the answers, it always seemed to match up quite well.

I: And did you also feel that the tone or urgency, the emotional content of the message was conveyed correctly, or did you feel that it came across differently somehow?

A: No, I think it came across as it should have.

I: Did you notice any differences in the use of medical terminology? Between the two methods.

A: Well, with ChatGPT, I noticed it more. Yes, I mean with Romanian, I only hear the medical terms because they sound very similar to Latin. But no, not really.

I: Okay, and if you noticed any inaccuracies or errors, how did that affect the consultation process?

A: I did not really notice that either. I just tried to ask again when something was unclear, and then I always got the answers I was hoping for. Actually, with both of them.

I: Okay. You already answered that you would prefer the interpreter if you were to do it again. But is there anything else you noticed that you would like to mention?

A: I always find it a bit difficult with interpreters, how to speak. Do you talk to the interpreter about the patient, or do you always address the patient directly? And I do not know how they translate it, so they will address him directly, but for you, which is better? Do I ask the interpreter about the patient, or do I just speak directly, as if I were really talking to the patient, and the interpreter is really just the medium? I am a bit unsure about that. Because with ChatGPT, I can really ask: "How are you?" With interpreters, I do not always know how to do that. I think I switch back and forth a lot.

I: Yes, that is actually very interesting. I can tell you that at university we do it as a kind of 'transmission medium' and we address the patients as if we were the doctors, and you as if we were the patients. But not all interpreters do that. Some just speak in the third person. And in my experience, for example, most doctors speak to me, rather than to the patients.

A: Yes, that is where the danger lies. Because I can only talk to her and get an answer, it is actually important that the patient also sees that I am actually talking to them. It is difficult, but um, I hope that was not too bad.

I: That was great. Thank you!

Abstract (English)

This thesis investigates the quality and perceived understanding of consecutive medical interpreting in comparison with Speech-to-Speech (S2S) translation technology based on a large language model (LLM), focusing on interactions mediated by ChatGPT and a human interpreter. Drawing on both the conduit and the triadic models of interpreting, the thesis examines how interactional dynamics influence communicative outcomes in a clinical context. A mixed-methods approach was adopted, combining quantitative and qualitative analyses.

Four simulated medical interactions were conducted involving two Romanian-speaking patients and one German-speaking physician. Each patient participated in two scenarios: one mediated by a professional human interpreter and one by ChatGPT. The consultations were transcribed, segmented, and annotated using three different frameworks. Understanding was quantitatively analysed by identifying evidence of understanding exhibited by the participants following Clark & Schaefer (1989), while quality was evaluated through a framework for quality assessment in medical interpreting proposed by Hamai et al. (2024) and the Multidimensional Quality Metrics (MQM). Understanding and perceived quality were further explored through semi-structured interviews, which were analysed qualitatively.

The findings indicate that while S2S translation technology constitutes an easily accessible and functional tool for information exchange, human interpreting generally results in higher perceived understanding and reliable quality. MQM results reveal that although ChatGPT can perform well, it remains less reliable, producing lower scores and critical errors compared to the consistent quality of the human interpreter. At the same time, the analysis based on Hamai et al. (2024) underscores the importance of the interactional capabilities of human interpreters and their contribution to effective communication in these scenarios. Furthermore, the results suggest that participants tend to modify their speech when interacting with ChatGPT, shortening their utterances and using standard language to facilitate machine processing. Lastly, the thesis addresses the ethical considerations of employing AI tools in contexts governed by strict confidentiality requirements.

The thesis highlights the need to evaluate interpreting quality and LLM-based S2S translation technology through both outcome-based measures and interactional dynamics, demonstrating the potential and limitations of such tools while underscoring the continued importance of human interpreters for accurate, nuanced, and ethically sound communication in healthcare.

Abstract (Deutsch)

Diese Arbeit untersucht die Dolmetschqualität sowie das subjektiv wahrgenommene Verständnis im konsekutiven medizinischen Dolmetschsetting und vergleicht diese mit einer auf einem großen Sprachmodell (LLM) basierenden Speech-to-Speech-Übersetzungstechnologie (S2S). Im Fokus stehen dabei Kommunikationssituationen im Kontext des Medizindolmetschens, die entweder durch einen/e professionellen/e Dolmetscher/in oder durch ein KI-gestütztes System (ChatGPT) vermittelt werden. Auf Grundlage des Conduit-Modells sowie des triadischen Interaktionsmodells des Dolmetschens untersucht die Arbeit, inwiefern die Interaktionsdynamik die kommunikativen Prozesse und die Dolmetschqualität in einem klinischen Setting beeinflussen. Methodologisch basiert die Studie auf einem Mixed-Methods-Design, das quantitative Auswertungsverfahren mit qualitativer Diskurs- bzw. Interaktionsanalyse kombiniert.

Es wurden vier simulierte medizinische Interaktionen durchgeführt, an denen zwei rumänischsprachige Patient:innen und eine deutschsprachige ärztliche Fachperson beteiligt waren. Jeder/e Patient/in nahm an zwei Szenarien teil: einem, das von einem/er professionellen Dolmetscher/in vermittelt wurde, und einem, das von ChatGPT vermittelt wurde. Die Konsultationen wurden transkribiert, segmentiert auf Basis von drei unterschiedlichen Annotations- bzw. Analyseframeworks systematisch kodiert und annotiert. Das Verständnis wurde quantitativ untersucht, indem Anzeichen für das Verständnis der Teilnehmer:innen gemäß Clark & Schaefer (1989) identifiziert wurden, während die Qualität anhand eines von Hamai et al. (2024) vorgeschlagenen Modells zur Qualitätsbewertung im medizinischen Dolmetschen sowie der Multidimensional Quality Metrics (MQM) bewertet wurde. Die Qualität und das wahrgenommene Verständnis wurden zusätzlich durch halbstrukturierte Interviews untersucht, die qualitativ analysiert wurden.

Die Ergebnisse zeigen, dass die S2S-Übersetzungstechnologie zwar ein leicht zugängliches und funktionales Instrument für den Informationsaustausch darstellt, das menschliche Dolmetschen jedoch im Allgemeinen zu einem höher wahrgenommenen Verständnis und einer zuverlässigen Qualität der Vermittlung führt. Die Auswertung mithilfe des MQM-Rasters verdeutlicht, dass ChatGPT solide Leistungen erzielen kann, allerdings weniger zuverlässig ist und, im Vergleich zur konstant hohen Qualität des/r Dolmetschers/in, niedrigere Bewertungen sowie kritische Fehler aufweist. Gleichzeitig unterstreicht die auf Hamai et al. (2024) basierende Analyse die Bedeutung der Interaktionskompetenz menschlicher Dolmetscher:innen und ihren Beitrag zu einer effektiven triadischen

Kommunikation in diesen Szenarien. Darüber hinaus deuten die Ergebnisse darauf hin, dass die Teilnehmenden ihre Sprache bei der Interaktion mit ChatGPT anpassen, indem sie Äußerungen verkürzen und Standardsprache verwenden, um die maschinelle Verarbeitung zu erleichtern. Schließlich thematisiert die Arbeit ethische Fragestellungen im Zusammenhang mit dem Einsatz von KI-Tools in Kontexten mit strikten Vertraulichkeitsanforderungen.

Die Arbeit unterstreicht, dass die Bewertung von Dolmetschleistungen und von auf großen Sprachmodellen basierender S2S-Übersetzungstechnologie sowohl über ergebnisorientierte Qualitätsindikatoren als auch unter Berücksichtigung der Interaktionsdynamik erfolgen muss. Dabei werden das Potenzial und die Grenzen solcher Technologien aufgezeigt, während zugleich die fortwährende Relevanz menschlicher Dolmetscher:innen für eine präzise, nuancierte und ethisch fundierte Kommunikation im Gesundheitswesen betont wird.