MASTERARBEIT | MASTER'S THESIS

Titel | Title

ChatGPT, DeepL or human translators: perceptions of translations from English into German in the Austrian Armed Forces

verfasst von | submitted by

Sophie Grossalber

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

Master of Arts (MA)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 070 331 342
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Translation Deutsch Englisch
Betreut von | Supervisor:	Univ.-Prof. Dragos Ioan Ciobanu PhD

Abstract

Aufgrund der prekären Sicherheitslage in Europa durch Russlands Angriffskrieg auf die Ukraine, sowie Desinformationskampagnen und der Weiterentwicklung von maschinellen Übersetzungstools ist es in militärischen Kreisen immer wichtiger zu wissen, woher Übersetzungen stammen und ob den Informationen in übersetzten Texten vertraut werden kann. Hierzu wird mithilfe der Multidimensional Quality Metrics von MQM untersucht, ob ChatGPT und DeepL in der Lage sind, militärische Texte in zufriedenstellender Qualität von Englisch auf Deutsch zu übersetzen. Gemeinsam mit einer menschlichen Übersetzung desselben Textausschnitts werden die rohen, maschinellen Übersetzungen fünf Teilnehmer*innen mit unterschiedlichem Erfahrungsschatz im Übersetzen und im Umgang mit maschinellen Übersetzungstools vorgelegt. Die Teilnehmer*innen geben anschließend ein Ranking ab, welcher Text ihrer Meinung nach aus welcher Quelle stammt und an welchen Kriterien sie diese Auswahl festmachen, um zu untersuchen, ob und wie Bedienstete im österreichischen Bundesheer zwischen maschineller und menschlicher Übersetzung entscheiden können und ob sie die Übersetzungen hinsichtlich Qualität unterschiedlich wahrnehmen. Sowohl in der vorangegangenen, manuellen Evaluation der Übersetzung als auch in den qualitativen Interviews zeigt sich, dass DeepL und ChatGPT zwar allgemeinsprachlich gute Übersetzungen produzieren, aber die militärische Terminologie ein Problem darstellt. Die Teilnehmer*innen können durch die Terminologie zwischen menschlicher und maschineller Übersetzung unterscheiden. Obwohl die Qualität der rohen, maschinellen Übersetzungen die Teilnehmer*innen überrascht, sprechen sie sich dafür aus, dass Post-Editing vor Veröffentlichung der Texte nötig ist. Weitere Studien mit einem größeren Teilnehmer*innen-Pool und anderen Sprachkombinationen werden empfohlen.

Acknowledgements

First, I want to thank my supervisor, Univ.-Prof. Dragos Ciobanu, PhD, for the support and the guidance throughout the year and a half that I worked on this thesis from the first idea of a subject to the final draft, and for the help in narrowing down my research field. I want to thank the Austrian Armed Forces' Language Institute, first and foremost Col. Mag. Mag. Thomas Fronek and ObstltdhmfD Mag. (FH) Dr. Karl Testor, as well as my study participants for the cooperation for this thesis. And last, but not least, I want to thank my family and friends for letting me ramble about my thesis when I got stuck and needed to find a way across a hurdle.

Table of Contents

Table of figures

1. Introduction and Previous Research

1.1. Motivation for Thesis

The debate around how to access ethical training data for machine translation (MT) and Large Language Models (Moniz & Parra Escartín 2023) as well as the move away from a solely physical battlefield with focus on maximizing gained territory to a hybrid one with calculated disinformation campaigns in social and traditional media, as evidenced in Russia's war against Ukraine, have increased the need for knowing whether a text and a translation originate from a trusted source. This need and the concerns around disinformation are reflected in current research as well as EU policies like the Digital Services Act, which came into effect on February 17th, 2024, and aims to "prevent illegal and harmful activities online and the spread of disinformation." (European Commission 2024) Previous research exists with a focus on the use of machine translation in the field (Center for Intelligent Information Retrieval 2024; Holland et al. 2000; Program Executive Office, Intelligence, Electronic Warfare, and Sensors (PEO IEW&S) 2016) and the relationship between interpreters and soldiers (Dijk 2021). This thesis aims to provide a starting point for further research into how evaluations of neural machine translations (NMT) and human translations differ among military personnel, as well as possible further training of personnel – both translators and soldiers.

The motivation for this research project does not solely lie in my personal interest in machine translation and the perception of target audiences of machine and human translation, as well as an honest concern of a professional future as a (freelance) translator given the "hype" of machine translation which have been debated countless times in different classes over the course of the MA programme. The research project is also informed by two internships I did with the Austrian Armed Forces' Language Institute in Vienna as part of the MA programme, which is where the desire for the chosen domain in defence originates. The three research questions are as follows:

- **RQ_1:** How well can DeepL and ChatGPT handle translating texts in the military domain from English into German?
- **RQ_2:** How do employees of the Austrian Armed Forces perceive and differentiate between translations done by ChatGPT, DeepL and human translations?

- **RQ3:** How does knowing a text is a machine translation change their perception of the translation?

1.2. Working definitions of machine and human translation

Since the focus of this thesis lies on the differences in perception between NMT texts and human translated texts, it is first necessary to define machine translation, neural machine translation and human translation and where these are located within the field of translation studies. Asscher argues in his article "The Position of Machine Translation in Translation Studies: A Definitional Perspective" that MT is inherently connected to the formerly popular prescriptive approaches of translation studies definitions, but it also "automatically" has a place within the newer descriptive approaches. However, Asscher also notes, we must remember that "translation studies has thrived in the absence of an accepted definition of translation" (2023: 4). This, by no means, insinuates that this thesis will operate without accepted definitions. Instead, the mention of Asscher's text is supposed to serve as a reminder that definitions in translation studies are not a rigid case of blindly following predecessors' arguments. Instead, they possess a long-standing tradition of fluidity, so to speak.

This thesis will use definitions of (neural) machine translation laid down in Koehn, Kenny and ISO 17100:2015 as guidelines for the necessary differentiation of translated texts' origins for the experiment reported on and analysed in later chapters. However, given the origin of this thesis within translation studies, definitions – or rather demarcations – given by participants of the experiment may differ from those widely accepted within translation studies as some of the participants may not possess a background in translation and prior, in-depth knowledge of the field. ISO 17100:2015 defines machine translation as "automated translation (2.1.2) of text or speech from one natural language (2.3.8) to another using a computer system" (2015), which not only Koehn and Kenny agree with but also the field of translation studies at large. It is to be assumed that the participants of this thesis' study will very likely agree with the definition laid out in the ISO norm above. However, this thesis does not aim to solely investigate how civilian and enlisted personnel in the Austrian Armed Forces define machine translation but rather how they will react to the knowledge that a given text has been translated by a machine system. Furthermore, Dorothy Kenny notes in chapter 2 of *Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence* that "translation, as practised by professional, human translators, requires the kind of intelligence that strong AI aspires to, but that such intelligence still remains beyond the

capacity of machine translation systems" (2022: 35). The "kind of intelligence that strong AI aspires to" talked about here encompasses the ability to refer to the context of both the translation brief and that of source and target text cultures among other things. Whether humans without prior translation training, such as perhaps some participants in this study, will be able to make those inferences when reading translated texts without knowing the texts' origins remains to be seen, as this thesis' focus does not include presenting the participants with a translation task but only the target text.

Apart from the before mentioned, still unattainable intelligence, Philipp Koehn mentions another problem for machine translation, whether that is neural machine translation or a statistical model. For him, the biggest challenge that natural language still poses is ambiguity, which should not present a problem for translations in specific domains with standardised terminology. However, the machine translation models used for this thesis were not specifically designed to handle terminology from the military domain and, in ChatGPT's case, were not even designed with translation in mind. According to Koehn, "[n]atural language is ambiguous on every level: word meaning, morphology, syntactic properties and roles, and relationships between different parts of a text." (2020: 5) This ambiguity already begins with the subtle, but still meaningful, differences between the German and Austrian locales in the German language. When selecting English as a source translation, DeepL, for example, does not allow the user in the free version to differentiate between the German and Austrian locales, which in turn could mean more effort needed by a post-editor to adapt certain terminology to the Austrian locale and the conventions set by the Austrian Armed Forces. One of the hypotheses for this thesis, in line with the research question on how Austrian Armed Forces' personnel perceive translations without knowing the origins, is that those with prior translation experience will pick up on possible differences in terminology between the human translation and the raw machine translation output from DeepL and ChatGPT.

With possible issues facing machine translation, and the ISO definition on what machine translation is, laid out, I will now briefly return to Asscher and his argument on machine translation's position within translation studies. Asscher notes that the descriptive definitional approach in translation studies as it is now widely adopted in the field is "compatible with both MT and human translation, just as a general theoretical framework would hope to be (Asscher 2022, 15)" (2023: 15). He further argues that there are good reasons to view machine translation as fully at home within translation studies, which I agree

with. Especially with the thesis's focus on translation perception. As mentioned above, machine translation is defined as "translation performed by a computer system" (ISO 2015) – sometimes specifically trained for this task, the specific domain it is employed in, and the language pairs needed. From this we can then infer that human translation is the translation of a text or part of speech from one natural language into another by a human, although it is still possible to use computer systems as evident with the use of CAT tools and Translation Memory systems. The use of those tools has no significant impact on the definition of human translation in my mind as the main task of translation is still performed by a human and not the machine translation system.

While machine translation systems are not yet ready to be fully "out-of-the-loop"-systems and require humans to post-edit the raw output as well as prepare and curate the training data, assuming that humans would not require proper training and revision before publication is false. ISO 17100 lists revision and the "two-man principle" as important steps in the translation process. However, because I cannot afford to pay a professional post-editor, the machine translation output used for the thesis will be raw. The raw machine output is also important as it would be likely that a professional post-editor would likely correct any terminology errors that would enable a comparison between the texts produced by DeepL and ChatGPT. It is to be assumed that the human translation will have gone through at least one iteration of revision before evaluation by this author and subsequent presentation to the study participants.

This chapter has mentioned neural machine translation as the type of machine translation used for the comparison to a human translation but has yet to explain what neural machine translation generally means. As defined above, machine translation is understood to be "automated translation (2.1.2) of text or speech from one natural language (2.3.8) to another using a computer system" (ISO 2015). *Neural* machine translation is thus machine translation using a system based on neural networks – which is where this specific approach receives the addition of "neural" from, to distinguish it from statistical or rule-based machine translation models that came before and use different approaches for building the translation model. Or as Koehn delves a little deeper into the schematics of a neural machine translation model: "The core components of neural machine translation are the encoder, which takes input words and converts them into a sequence of contextualized representations, and the decoder, which generates an output sequence of words." (2020: 133) A neural machine translation model is thus made up of different parts that produce the, still raw, machine output

– also known as the translation. While DeepL is open about the technology and build of the model behind their translator – as mentioned, it is also a machine translation model based on an artificial neural network –, OpenAI is not as forthcoming about the schematics of their ChatGPT model. What is clear is that there is also a transformer component involved in OpenAI's Large Language Model. Fully understanding the schematics of ChatGPT or not, however, should not hinder the understanding that ChatGPT was not designed with translation as the main task it should perform but DeepL Translator was. Although both DeepL and ChatGPT are grounded in artificial intelligence and work with artificial neural networks.

1.3. Translation evaluation

Due to the nature of ChatGPT's main tasks, the however small difference to DeepL Translator and the fact that this thesis also investigates whether untrained Armed Forces personnel do or do not differentiate between machine and human translation, it is necessary to carefully select the approach for evaluation of all three target texts. Daems and Macken note (H)TER (Translation Error Rate) as the most often used automatic metric for the evaluation of MT as it "quantifies the amount of editing (insertions, deletions and substitutions of single words as well as shifts of word sequences) applied to the translation suggestion to create the final translation" (2020: 56–57). As the evaluation of the target texts by me solely serves to create a backdrop for the comparison of participants' answers, an automatic evaluation method did not seem suitable at the time of planning the thesis. Daems and Macken did, however, also argue that "readers and translators are not necessarily capable of distinguishing between human-translated and machine-translated texts (Vasconcellos and Bostad 1992; He et al. 2010)" (2020: 50). They went on to pose the question of whether revisers and post-editors would be able to make this distinction, but that shall not be part of the research area for this thesis, which will focus on the target audience's perception and ability to differentiate between human and machine translation. Daems and Macken's argument is still reflected in the hypothesis adjacent to the second research question in that it is to be assumed that Austrian Armed Forces' personnel will not be able to differentiate between the two types of translations without having received the proper training first. A view which would also align with my own prior experience with trying to distinguish human and machine translation from each other. Daems and Macken came to the conclusion after conducting their study on post-editing and revising human and machine translations that

> participants performed better when they had the wrong assumption about the provenance of a text (post-editing human translations and revising MT), which indicates that assumptions

> about the nature of a text are likely to influence the optimality of the revision or post-editing process. (2020: 69)

It is precisely these comments on readers and translators' (in-)ability to differentiate between human and machine translation and the change in performance when assuming the wrong origin of a text that has led to the decision for the first part of the thesis's experiment to not tell participants which chunk of a translation they are being presented with was done by a human or one of the two computer systems chosen. This is primarily to investigate whether they could tell which one was translated by ChatGPT, DeepL Translator or the human translator chosen but also to not influence their perception of the target texts before they had a chance to read them but rather for them to able to "go in blind".

As already mentioned above, an automatic MT evaluation method is unsuitable for this thesis's scope and research aim because the three different target texts necessitate a more general, adaptable to both machine and human translation errors, framework for evaluation, as does the later attempt to model answers from study participants to said evaluation framework. Thus, the choice fell on the Multidimensional Quality Metrics system developed in the QTLaunchPad project funded by the European Union. Since its development, the MQM framework has gone through a revision, resulting in MQM 2.0, which this thesis will use for translation quality evaluation to provide a most accurate basis for evaluation of the target texts, as well as the desire to use a framework that enables the attempt to model study participants' interview answers to. Time constraints and a lack of resources have already been mentioned in passing, thus it seems repetitive to mention them again. However, they did play a role in the decision-making process on which evaluation framework to use for the evaluation of translation quality of the investigated translations. Lommel et al. describe a hierarchical list of issue types which were "derived in a careful examination of existing quality evaluation metrics and the issues found by automatic quality checking tools" (2014: 458), while leaving out metrics for the assessment of project management and project quality and focussing solely on translation quality evaluation.

The MQM framework in its revised version contains seven parent categories for translation errors, with some of them also providing child categories for a more in-depth evaluation. Since this thesis's study participants might no have prior experience with translation, let alone translation evaluation as it is laid out in the MQM, and it seems unnecessary for the short source and target texts to expand the framework, we will stick to the

core MQM error typology. As defined on the MQM's website, this leaves the following parent and child categories for error typology:

- **Terminology**
  - Inconsistent with terminology resource
  - Inconsistent use of terminology
  - Wrong term
- **Accuracy**
  - Mistranslation
  - Overtranslation
  - Undertranslation
  - Addition
  - Omission
  - Do not translate
  - Untranslated
- **Linguistic conventions**
  - Grammar
  - Punctuation
  - Spelling
  - Unintelligible
  - Character encoding
  - Textual conventions
- **Style**
  - Organization style
  - Third-party style
  - Inconsistent with external reference
  - Language register
  - Awkward style
  - Unidiomatic style
  - Inconsistent style
- **Locale conventions**
  - Number format
  - Currency format
  - Measurement format

    - Time format
    - Date format
    - Address format
    - Telephone format
    - Shortcut key
- **Audience appropriateness**
    - Culture-specific reference
    - Offensive
- **Design and markup** (MQM Council 2024)

Due to the chosen source text chunk being only a small part of NATO's informational subpage on the Comprehensive Assistance Package for Ukraine and design and markup do not matter too much with what the study is trying to investigate, it was decided to leave out the parent category "design and markup" in its entirety, as well as some of the subcategories like "character encoding" from "linguistic conventions". Categories in the error typology that will not be considered have been marked through underlining and will again be mentioned in the second chapter. In this chapter, we shall only briefly take a more in-depth look at one of the parent categories as it relates to the specific needs and intended addressees of the source and target text used in this thesis.

"Audience appropriateness" was called "Verity" in the first version of MQM and, although the subcategories within it do not seem to be necessarily relevant to the target audience of the texts used in this study, the concept of "audience appropriateness" as a whole does very much so. Lommel et al. described "Verity" as a concept that "refers to extra-linguistic truth correspondence in texts" (2014: 459). So, the category is supposed to catch issues or errors that would arise in texts used in highly specific contexts and cultures. According to Lommel "audience aapropriateness/Verity" goes beyond simple errors that could be found while evaluating the target text and looking at the source text, and instead "requires the evaluator to consider the text in its intended environment" (2014: 459).

Judging the translation for how well it fits into its intended environment and is appropriate for the target audience becomes especially important for translations within the military domain – second probably only to terminological accuracy and linguistic conventions found in the target language. Errors in appropriateness and accuracy can severely impede the understanding of and trust in a source text within the military domain. Alternatively, it could be exactly those errors could be used to discredit actually trustworthy sources. While, for

example, typography errors should not appear in official communication – even if that communication takes place on a social media platform, or rather, especially then –, errors can still be made and thus deter from the intended message.

1.4. (Machine) translation in the military

(Machine) Translation systems are nothing new within the military domain as is evidenced not only in the various embedded systems developed by the U.S. military like SYSTRAN, which began as a U.S. Air Force programme in 1968 (SYSTRAN by ChapsVision 2024), or FALCon (not to be confused with a DARPA/U.S. Air Force project for rocket development, sometimes spelled "Falcon"), GALE, MFLTS and LORELEI – all systems developed by DARPA, the U.S. Defence Advanced Research Projects Agency, and which include embedded MT. "Embedded MT" refers to machine translation systems that are embedded in a bigger software complex with machine translation only as one stage of the process (Holland et al. 2000). However, while it seems unlikely that the U.S. would be the only country to be actively researching machine translation and embedded MT for applications in a military environment or in the field, information on DARPA projects is still easier to come by than other countries. Although searches do not yield a lot of results. Bart Maczynski, VP of Language Weaver, mentioned in an interview for the SlatorPod podcast that Language Weaver Edge was deployed in a project with the U.S. military in cooperation with South Korea (Slator 2024: 5'40"-6'10"). During research for the thesis, I also found a registered patent from 2020 in the European Patent Office's database for a machine translation optimisation method and system from China, which "aims to provide a machine translation optimization method and system in the national defense war industry field so as to ensure translation consistency of keywords in a whole article and improve translation quality" (Yao et al. 2020: 1). The U.S. seems to simply be the country willing to talk about their projects the most.

This is also evident in an article by Holland et al. from 2000 which reports on three separate military field exercises during FALCon's development which included an evaluation of the system by the end users – who did receive training to use the system during those exercises as well. Sadly, this article is the only one found that details and analyses the evaluation of the system by the intended end users. This section will thus focus on FALCon and the article detailing the evaluation by its end users. Another paper which shall be talked about more in-depth later in this chapter, deals with the investigation of the importance of human interpreters for international military operations. According to Holland et al., FALCon

was “a hardware-software prototype developed for the U.S. army that permits paper documents in selected languages to be scanned, processed by optical character recognition (OCR), and automatically translated into English ([7])” (2000: 240). FALCon was deployed in field exercises in Bosnia, Haiti and Central America between 1997 and 1999 with the processed documents ranging from memos from coalition partners of those operations to publicly available materials like newspapers, and the source languages included Arabic, Croatian, Russian, Serbian and Spanish.

One of the critique points brought up with FALCon was “users’ judgement that manual OCR correction took so long, it rendered ‘the current system...less than desirable for the intended function.’” (Holland et al. 2000: 242). Nevertheless, participants in the FALCon project’s evaluation insisted that the concept still possessed validity, as is also evident in the continued development of machine translation for the U.S. military in various DARPA projects. The date of development and the year of the evaluation should also be of note, however. FALCon was developed and deployed between 1997 and 1999, more than 25 years ago. OCR systems have improved in that quarter of a century – as have machine translation systems. Nevertheless, the article provides valuable insight into the needs of end users of (embedded) machine translation systems within the military domains, their expectations towards the output quality and possible use cases for (embedded) machine translation.

FALCon was meant to provide an automated possibility for screening documents for their relevance to the mission and its staff before those documents are handed over to professional translators and linguistic experts. Thus, FALCon was never intended for full-document translation with human-like quality but rather so-called gist translation. As mentioned above, the OCR capabilities of the FALCon system appeared to be the biggest dampener in its applicability during the field exercises. Noise originating from low-quality documents and images posed a challenge as “noise is not predictable from a small set of documents used to test performance in the lab” (Holland et al. 2000: 243). The experiment conducted for this thesis could also be seen as an in-field exercise, similar to the exercises conducted during FALCon’s development and evaluation, as the participants might be enlisted personnel and the target audience of the translated text. As has already been mentioned, participants will be presented with already translated text, divided into equal sections in line with the paragraphs in the texts, and asked to rank the chunks from least likely to be machine translation to most likely to be machine translation. Afterwards, they will be asked to justify their choices, hopefully giving answers akin to the core MQM error typology used to evaluate

the translations beforehand. Finally, the study participants will be informed which text was translated by DeepL, ChatGPT or the human translator and asked whether that information changes their perception of the texts.

Thus, while not embedded in a specific software environment, the machine and human translations done as part of the experiment could still be seen as embedded translation systems or processes as they are part of the experiment process as a whole. While FALCon's evaluation ended up being predominantly positive, possibly in regards to the potential of the system's future application and not the satisfaction with the results at the time of the evaluation, one critique point – apart from the difficulties with the OCR component – was crashes experienced by users "every time they tried an Arabic translation, crashes that did not appear during software testing in the lab" (Holland et al. 2000: 244). Those incidents experienced in the field could be detrimental to the mission and to people's lives were to occur in serious scenarios and environments. One could take this caveat to use as an argument for human interpreters and translators and against the use of machine translation in the field, however, it is important to note that FALCon – as well as any of the other DARPA research projects – does not seem to have been intended to fully replace human experts as is often proclaimed when machine translation is brought up in a conversation these days, but to be used as a tool to lessen the potential workload for the human experts and thus streamline the processes and enable the experts to focus on the important documents and intelligence. The same is true for this thesis. Machine translation is to be seen as a tool capable of aiding translators and end users and not as a replacement of human personnel.

This sentiment is also echoed in Andrea van Dijk's thesis *Talk up Front: The Influence of Language Matters on International Military Missions with a Particular Focus on the Cooperation between Soldiers and Interpreters* as it investigates and makes a case for the importance of either trained linguistic experts to be used in the field or local experts for the success of an international mission. Van Dijk also argues that innovations made in the recent years within machine translation research and development "will never replace the interpreter as an intermediate between parties. Like war itself, interpreting is and will remain a human and social activity." (2021: 10) As can be seen throughout this chapter, experiments with machine translation and applications in the field have been conducted before, and van Dijk does not argue fully against deploying machine translation in the defence sector, simply for recognising that human actors in the mediation process or on missions are still more capable and more important to the success of these missions than the machines. War is still fought by

people, however big the shift to a hybrid or fully cyber battlefield, as seen not only with Russia's war against Ukraine but also with the situation in Israel and Gaza.

Van Dijk notes the importance of interpreters, how dependent soldiers are on them and how they are not awarded the acknowledgement of the vital work they are doing, stating that "[a]s a consequence, interpreters remain relative outsiders to the military organization" (2021: 11). Whether that is due to a lack of awareness of their importance to a successful mission or due to them not being trusted enough as they could be seen as being caught in the middle between the military organisation that employs them and, possibly, their country and culture of origin, is unknown. Van Dijk focusses on the relationship between soldiers and interpreters, but this thesis does not, as should be clear by now. However, the lack of acknowledgement and possible lack of awareness of translators' and interpreters' work is likely to come up in the experiment conducted as part of this thesis.

For the successful transfer of meaning from the source language into the target language, van Dijk argues that an interpreter's prerequisite is the ability to command "a solid and sound understanding of the subject matter of the military operation and/or negotiation" (2021: 23). DeepL and ChatGPT are not necessarily familiar with these topics. While ChatGPT might have accidentally acquired context and necessary weights to be able to predict the correct terminology from a given prompt and text, it seems highly likely the publicly available model of the system was not specifically trained on the terminology. The same seems to be true for DeepL. Although governments and international institutions do also provide open-access corpora with aligned texts in various language pairs that can be used to train a machine translation model, web-crawled and controlled corpora might be even more useful for terminology training, if there is no official, normed terminology available. Much like human translators and interpreters require training to understand the intricacies of the domains they specialise in, so do soldiers and machine translation systems. Soldiers, as the target audience for translations, need to learn the terminology and the intricacies as well, but also need to command a certain level of understanding of the source culture and language if they should be able to evaluate translations – and the author of this thesis argues that a heightened awareness of what constitutes translation and what to look out for with machine translation should strengthen their country's and society's resilience to external attacks through, say, social media. Disinformation, the threats posed by it and the potential for misuse of machine translation to further disseminate mis- or disinformation will be discussed further in the next section.

The U.S. Department of Defence also evaluated possible lessons learned from their operations in Iraq and Afghanistan and consequently shifted their doctrine. “Language learning and cultural awareness have since become qualified as priority matters that need to be taken up in future policies of military effectiveness (Moskos, 2007, 3-13; McFate & Jackson, 2006, 14).” (Dijk 2021: 21) Whether future policies were updated accordingly is not part of this thesis and shall not be discussed.

The intended audience providing feedback to the linguistic experts – whether those be human or machine – is also an important part of training future linguistic experts. This chapter has already mentioned training data for machine translation and the fact that clients and target audiences rarely provide feedback to the translator, unless specifically asked for feedback. However, one could argue, target audiences and clients are not equipped to provide feedback on things such as linguistic conventions and whether a translation reads idiomatically or not. For specialised domains such as the military and the defence sector, a lot of the terminology and some phrases are standardised to enable easier communication and English – for international operations – is still the lingua franca, unless another language presents itself as a better choice in the context of specific missions. Translators “embedded” in the military network, i. e. either translators who are also enlisted or civilian employees of the Armed Forces, do have access to this terminology and do receive feedback when time and setting allow. During my internships with the Austrian Armed Forces as part of my studies, I did have access to various glossaries and was involved in terminology research for cases where there were no glossaries to access at that moment. The access to terminology resources and experience were reasons for my asking one of the current translators with the Austrian Armed Forces to contribute to the experiment by translating the chosen source text which will then be used as a sort of “gold standard” to compare the machine translations to. Especially in terms of terminology, this should enable me to provide an accurate assessment of errors and the accuracy of ChatGPT and DeepL during the first evaluation process in the experiment scenario.

Another issue for translators and military personnel without linguistic and cultural training of the country and region they are deploying to was brought up by van Dijk when talking about the former UN leader General Roméo Dallaire and his autobiography in which he recounts his time in Rwanda during the civil war. Van Dijk describes Dallaire’s account as “is exemplary for the fact that a proficiency of the English language, the proclaimed international military lingua franca, is insufficient for facing the socio-linguistic challenges of

an environment dominated by a multitude of vernaculars" (2021: 17). Standardised terminology can be taught and learned – to humans and machines – but vernaculars and accents are little more difficult. Nevertheless, with enough training data and time to train machines and human experts, those vernaculars and accents become less of a challenge. If there are enough resources available for a language as some lesser-known languages are not present in current machine translation systems as options because there is not enough training data to support training the systems in those languages as well.

1.5. Machine translation and ethics

Within developing machine translation systems, training of MT systems and actively using them, "we should also consider the carbon emissions produced when developing MT systems, and when using them in production" (Moniz & Parra Escartín 2023: 3). Now, the biggest part in this thesis's experiment is the comparison of machine and human translation output. While I might not know exactly how much CO_2 I produced with having DeepL and ChatGPT translate the source text, I am aware that those emissions are not going to be zero. The focus of this thesis is not to investigate whether the participants are aware of the amount of carbon emissions produced during development and use of machine translation systems, however, it is possible for participants to bring up their own awareness.

They might also be aware of the implicit and explicit biases in the training data that "need to be addressed to ensure that the output of the MT systems is not biased and hence cannot potentially harm a particular societal group (e.g. an MT system that produces translations that are sexist, racist, etc.)" (Moniz & Parra Escartín 2023: 3). Moniz and Parra Escartín call this "Responsible MT", a system designed ethically and in such a way that "data bias, data licences and rights, ecological footprint, and intended end-users" (2023: 3) have been considered before development and training of the system began. But finding the biases in the training data – and thus the biases still present in the machine output – also requires awareness on the end users and evaluators' parts. Which still necessitates giving end users and evaluators the correct training to do so, which in turn, is costly and time-consuming. This thesis shall, however, provide a starting point for further training of both target audience, so soldiers and personnel within the Austrian Armed Forces, and the translators and linguistic experts working for them. It is by no means intended to be a fully-fledged study, but rather a pilot, designed to hopefully point out the starting points for future training, and perhaps even development of a machine translation system specifically designed for the deployment in the

field of defence. The latter simply requires consideration of the points mentioned above, specifically biases and environmental impacts.

1.6. Threats posed by machine translation – disinformation and propaganda

As mentioned on the first page of this chapter, disinformation is a current issue of concern for not only the Austrian Armed Forces or NATO, but also the EU and citizens in general. Whether that be propaganda or fake news, mis- and disinformation can and does threaten countries and people's lives. Whether that is an ill-intentioned rumour shared by a colleague or troll and bot farms swarming the replies on a post in a social media network as they are employed by Russia to attack Ukrainians and their allies and undermine morale (Garrigoux 2022). Although I could not find any research on the use of machine translation in such scenarios, it seems unlikely that bad actors would pay translators to translate their likely vast amounts of ill-intentioned messages into the target language and machine translation would thus be the natural modus operandi as it would be cheaper. It could also be argued that any typos or awkward phrasings present in the items of dis- and misinformation could be placed or left there deliberately to lend a certain degree of credibility to the bots' legitimacy as "real people".

First, let us define and differentiate between mis- and disinformation as they are often used synonymously but are not actually synonyms of each other. The Council of Europe defines misinformation as "false information shared with no intention of causing harm" (Council of Europe 2024: 5), for example, a post full of false information shared by a friend because it was shared by friend of theirs or another, bigger account from a person they trust. There was no ill-intention behind the action of sharing the information, however, it may still cause significant harm. Disinformation, on the other hand, is defined as "false information shared intentionally to cause harm" (Council of Europe 2024: 5), so, for example, a reply formulated by a troll or posted by a bot. The *Democratic Schools for All* site from the Council of Europe goes further than a simple definition of mis- and disinformation in that it argues that "[t]he ability to respond critically to online propaganda, misinformation and fake news is more than a safe-guarding tool, however, it is also an important democratic competence in its own right" (2024: 6). Which brings to mind again what has been said before about how the awareness of what to look for in social media posts or websites or translations received by the target audience to be able to spot non-native texts, i. e. machine translated, false information

designed to confuse or distract from another cause and real information posted, for example, by an official institution.

Audrey Garrigoux argues that "[a]lthough the use of disinformation as a weapon has always existed, the social media landscape has multiplied its reach and potential penetration" (2022: 1). We can also assume that machine translation and openly accessible, commercial machine translation systems like DeepL (or even ChatGPT) even with their restrictions on file size and source text length have also contributed to the growing threat of disinformation being used as an effective weapon off the physical battlefield. Garrigoux's article focusses on disinformation and its connection to Russia's war of aggression against Ukraine and not only provides numbers but also more details on the platforms and the kind of disinformation spread by Russian actors in an attempt to undermine Ukrainian morale and influence the narrative of the war outside of Russia:

> Efforts to manipulate public opinion on social media took place on Twitter and Facebook, with extensive efforts also concentrated on Instagram, YouTube and TikTok. Evidence also exists of disinformation campaigns taking place in the comments sections of major media outlets (The Guardian, 2022[15]). (Garrigoux 2022: 3)

According to Garrigoux, who cited Facebook as an example, more than 50 countries were being targeted by foreign disinformation operations. Those operations were not necessarily conducted by Russia, however, Facebook's report, according to Garrigoux, also cites the United States, Ukraine and Great Britain as the most frequently targeted countries. That, by no means, is intended to imply that only those three countries are being threatened by disinformation campaigns that are solely conducted by Russia. But, due to the urgency of the war against Ukraine, Ukraine's geographical proximity to Austria and Austria's position as an apparent base of operations for Russian spies[1], the U.S., Ukraine and Great Britain are not the only countries threatened and for whom it is necessary to take steps and implement measures to improve resilience against disinformation and what is known as psychological warfare.

In fact, Russia's disinformation campaigns apparently go far beyond influencing the United States elections and attacking morale in Ukraine as Garrigoux points out that "[…] in Africa, Russian and Russian-affiliated actors have created messages disparaging democracy and spreading misleading and false information about political actors, as well as narratives seeking to stoke social tensions (Africa Center for Strategic Studies, 2022[61])." (2022: 10)

[1] See, for example, Simoner, Michael (4 April 2024). "Spionieren für Russland hat in Österreich Tradition". Wien: *Der Standard*. https://www.derstandard.at/adblockwall/story/3000000214480/spionieren-fuer-russland-hat-in-oesterreich-tradition [last accessed 24 June 2024]

Again, the sheer volume of text required for their campaigns and operations very strongly suggests, Russia and other bad actors either are paying a lot of translators or they are using machine translation to achieve their goals and fulfil the desired volume.

Thompson et al.'s investigation also points towards the use of machine translation in disinformation campaigns. They looked at a multitude of translated texts on the web and came to the conclusion that " […] the quality of these multi-way translations indicates they were primarily created using M." (2024: 1). Their article also raises concern about the quality of training data for machine translation and large language models if said training data sets are crawled from the web as they could have been either generated by a model like ChatGPT or machine translated without significant post-editing. Thompson et al.'s investigation indirectly supports the suggestion that bad actors could use machine translation to conduct their disinformation campaigns in a variety of languages across various regions and social media platforms.

On the threat posed by disinformation towards the EU, the European Court of Auditors conducted an audit into the EU's action plan against disinformation, which was presented in 2018 and found one of its expressions in the Digital Services Act which came into effect on February 17th, 2024. The audit covered the period from 2018 until September 2020 and in it, the European Court of Auditors concluded that "[…] the EU action plan was relevant but incomplete, and even though its implementation is broadly on track and there is evidence of positive developments, some results have not been delivered as intended" (European Court of Auditors 2021: 4). However, the audit was published in 2021, so it would be interesting to know what the results would be today, now that the Digital Services Act has actually come into effect. Quantifying the effects of the Act, however, likely will take a couple years.

The audit also included an investigation into a framework for engagement with online platforms provided by the European Commission, however, the European Court of Auditors found that "[…] the code of practice fell short of its goal of holding online platforms accountable for their actions and their role in actively tackling disinformation" (European Court of Auditors 2021: 5). This code of practice does highlight another issue in the fight against disinformation, apart from implementing policies and guidelines for media and AI literacy training for civilians and governmental institutions' personnel: governments cannot actively combat disinformation on social media or in the comment sections of traditional media's websites without infringing on the fundamental rights of freedom of opinion and

expression. Unless the social media platforms implement effective strategies to combat disinformation, trolls and bots, governments might see banning respective platforms as the only way to combat disinformation apart from previously suggested policies for media literacy training and laws which might infringe on the already mentioned rights[2]. This would likely also include any legislation regulating the development and use of Large Language Models and machine translation systems intended for commercial use.

Four years ago, Sofia Martins Geraldes argued in her chapter in Moehlecke de Baseggio et al.'s book *Social Media and the Armed Forces*, that "[…] the debate on social cyber-attacks and the hostile use of social media has been centred on its use in peace-time situations, focussing mostly on its effectiveness as a tool to disseminate disinformation […]" (2020: 190). Four years later, I would argue that the debate has definitely shifted its focus to the use of social media for hostile activities in peace times to the use of social media as an active cyber-weapon in an active war zone as not only Russia and Ukraine have shown, but as is also evident in the situation between Israel and Palestine. Sofia Martins Geraldes even points the shift out herself in the chapter before Russia's full-scale invasion of Ukraine: "The Russia-Ukraine conflict demonstrates the growing influence of cyberspace and social media as a digital platform in the conflict landscape of the twenty-first century in support of conventional military action." (2020: 190)

Martins Geraldes' chapter also goes into more detail on how Russia and other bad actors use disinformation campaigns and social media platforms' inherent functions and mechanisms to undermine governmental institutions, also by "describing [the Ukrainian government] as corrupt and fascist" (2020: 200), which resulted in mistrust in the government, the Ukrainian military and its operations as well as a drop in morale and desertion. Due to the linguistic situation in Ukraine, with Ukrainian and Russian as widely spoken languages however, a deployment of machine translation in this scenario seems unnecessary and unlikely. But, as has been mentioned a couple of times now and as the author of this thesis will keep arguing, it seems highly likely that machine translation was and probably is still being deployed in disinformation campaigns by Russia when the area of

[2] See, for example, Paul, Kari (24 April 2024). "Senate passes bill banning TikTok if parent company does not sell it". *The Guardian*. London: Guardian News & Media Limited. https://www.theguardian.com/technology/2024/apr/23/senate-passes-bill-banning-tiktok-will-parent-company-sell [last accessed 24 June 2024]

operation is not focussed on Ukraine but other countries – particularly in Europe and the EU's Western allies.

In the chapter following that of Martin Geraldes in the same book by Moehlecke de Baseggio et al., Norri-Sederholm et al. investigate and detail the threat faced by Finland through social media and disinformation campaigns, but the authors also mention the strategies employed by the Finnish government to mitigate those threats – one of those being generalised trust and the before-mentioned media literacy training. The authors go on to argue that "democratic countries, such as Finland, cannot weaponise social media, as it is simply not acceptable to use these kinds of mechanisms and acts" (2020: 213). Thus, these democratic countries – and Austria is one of them – need to devise and employ other strategies. The EU is clearly attempting to offer guidelines and legal frameworks with policies and Acts such as the Digital Services Act and publications on literacy media training like the one from Democratic Schools for All, yet it remains the responsibility of the member states' governments to implement relevant policies to provide the necessary training.

As Norri-Sederholm et al. mention in their chapter about social resilience and media literacy training

> […], measures to promote media and information literacy can help citizens navigate the social media environment. In fact, research suggests that building social resilience is also an essential contributor to resistance towards disinformation and misinformation (City of Helsinki 2018; EEAS 2019a; West 2017). Social resilience refers to a society's ability to cope with uncertainty and to "recover from shocks and emergencies" (Giacometti et al. 2018, 5). (2020: 214)

Most of the literature cited in this section about the threats to the EU, its member states and Austria in particular, does not mention machine translation in their texts, however the comments accompanying the citations should have made clear that investigating the connections between machine translation and disinformation campaigns is a worthy endeavour and should be taken on in the future.

1.7. Outlook on following chapters

To quickly recap, the first chapter of this thesis is meant to introduce to the topic of the thesis and the motivation for undertaking this research project. This chapter has laid out the definitions of machine and human translation, briefly touched upon the research questions with some hypotheses already addressed and the methodology employed within the experiment that is part of this thesis. The chapter has outlined definitions for disinformation,

previous research done in the field of translation and interpreting in the field and argued for the importance of viewing machine translation as a tool for translators, interpreters and the target audience, instead of aiming to replace the human component in the translation process with a machine. The chapter also briefly touched upon Responsible MT and the economical and ethical concerns now attached to machine translation research, development and use. And there has been a brief discussion of strategies employed to bolster resilience to disinformation campaigns and how Russia uses disinformation in its war against Ukraine as a cyber-weapon.

The following chapters will, in order, focus more in-depth on the research questions, hypotheses and a detailed explanation of the methodology as well as a brief analysis of the source text. Chapter three will give an overview of the answers as given by participants in the study as well as an analysis of the results and an evaluation by the author of the thesis of the target texts. The final chapter will lay out the conclusions drawn from the experiment, followed by the bibliography and relevant excerpts from the interview transcripts.

2. Research Questions, Methodology and Source Text analysis

2.1. Research Questions and Hypotheses

As previously mentioned, this thesis aims to investigate the perception of translations by Austrian Armed Forces' personnel (also referred to as "target audience") by conducting a limited qualitative study. The research questions, which have already been briefly presented in Chapter 1, are as follows:

- **RQ_1:** How well can DeepL and ChatGPT handle translating texts in the military domain from English into German?
- **RQ_2:** How do employees of the Austrian Armed Forces perceive and differentiate between translations done by ChatGPT, DeepL and human translations?
- **RQ_3:** How does knowing a text is a machine translation change their perception of the translation?

The research questions are ordered hierarchically according to granularity with RQ1 being the broader, over-arching question to be answered and RQ2 following close behind with a specific focus on the Austrian target audience and their reception. RQ3 serves as a follow-up to RQ2 as necessitated by the experiment set-up: the study participants will not be informed, at first, which of the texts they are asked to rank were created with ChatGPT, DeepL or by the human translator employed by the AAF. The original "target audience" category included mainly untrained personnel with no prior experience with the translation process except as clients tasking the translators with a translation. However, due to a lack of volunteers, the experiment parameters were changed to employees of the Austrian Armed Forces' Language Institute, some of which have prior translation experience, and some do not.

The above research questions and the previous research as laid out in Chapter 1 lead to the following hypotheses, ordered by the same level as their corresponding research question:

- **H_1:** DeepL can produce a fairly fluent target text in German but will probably mix terminology from the German Bundeswehr with the Austrian Armed Forces and the German locale with the Austrian one, while ChatGPT might be able to disguise terminology issues and distract with issues in style and idiomatic use of language.
- **H_2:** Austrian Armed Forces' employees with no prior experience with ChatGPT, DeepL or translation in general will not be able fully differentiate between machine and human translation but might catch some of the more obvious errors from the

machine translations. While employees with prior translation experience might be able to catch the machine translation or LLM translation more quickly. Both categories of people might not initially note any distinctions in perception of the target texts beyond surface level errors.

- **H_3:** After being told of the origins of the target texts, respondents might make changes to their ranking and note differences in their perception of the texts.

A more detailed analysis and evaluation of the machine translation and LLM translation will follow in Chapter 3 before the presentation of interview results. H1 is primarily based on previous experiences with DeepL and ChatGPT, but also grounded in literature. As Kenny notes, a translation produced by a human will always be different and result in a slightly different target text; she does argue that "maybe we should allow machine translations to do the same" (2022: 32), however, a slight change in meaning for a target text in the military domain could result in catastrophic consequences in the reception and for a potential mission. Thus, I argue that this "allowance" for a shift in meaning in a machine translated text should be carefully considered within the context of the domain and the intended function of the target text.

The working definitions that will be used in this thesis have already been explained in Chapter 1, however, to recap, the thesis works with the following definitions:

- Machine translation is the automated translation of a text or a part of speech from one natural language into another by a computer system, i. e., ChatGPT or DeepL, which are used for the experiment in this thesis
- Human translation is the translation of a text or a part of speech from one natural language into another by a human (use of CAT tools and Translation Memories included) (ISO 2015)

This chapter will not go into detail on where to position machine translation within the field of translation studies, as that has already been done in Chapter 1. Criticisms of machine translation from researchers within the field have also been mentioned in detail in the previous chapter. The machine translations used in the experiment will be evaluated in detail in Chapter 3 to also provide a backdrop for the comparison between participants' answers and the score cards.

As is evident from the first pages of this chapter, the research questions are closely linked to each other and the overall topic of the thesis, which in turn was informed by the

author's internships with the Austrian Armed Forces' Language Institute as part of her studies. The aim for the thesis is not to point a metaphorical finger and prescriptively tell which machine translation system is better or whether human translators are still the gold standard, but to investigate if and how the target audience's perception differentiates between human and machine translation. Due to time and resource constraints, this thesis should be seen as a pilot study, enabling further study of the topic and different language pairs at a later stage, and to, perhaps, provide a starting point for training of both linguistic and non-linguistic personnel to raise awareness of machine translation and the issues still surrounding it.

2.2. Methodology

2.2.1. Translation Quality Evaluation

As has been briefly touched upon in Chapter 1 and above in Chapter 2.1., this thesis involves an evaluation of two machine translated texts with the use of the Multidimensional Quality Metrics (MQM). To give the most accurate evaluation, the decision was made to go with MQM 2.0 as the old version of the error typology has now become obsolete and categories like "Audience appropriateness" seem more fitting for the purpose of the experiment and the translations than its predecessor titled "Verity".

The reason for choosing the MQM framework was, on the one hand, familiarity with the framework through the author's studies at the University of Vienna and, on the other hand, the framework's inherent multi-purpose function. As RQ_2 notes, the experiment will attempt to focus on the differences in perception from the target audience of human and machine translation. MQM was built to cater to evaluators of both human and machine translation and thus lends itself nicely to the purpose of the thesis' experiment. For this purpose, a decision was made to use the core MQM error typology with the seven main categories being **Terminology**, **Accuracy**, **Linguistic conventions**, **Style**, **Locale conventions**, **Audience Appropriateness** and **Design and markup** during evaluation of the machine translations. Due to the nature of the source text as a website text and the intended focus on linguistic differences, it was decided to leave out Design and markup as a whole as the piece of the source text used in the experiment will be devoid of markups and ChatGPT and the web version of DeepL translator do not adhere to the original formatting either.

The scorecard was created in Excel with the formulas and design based on the documentation from the MQM Council. The main error categories, as seen in Figure 1, are

foldable and reveal the corresponding error subtypes. Their parent categories sum up all errors in that category, which are then multiplied with the Severity Penalty Multiplier and the corresponding Error Type Weight to calculate the Error Type Penalty Total (ETPT). The ETPT is the basis for the Absolute Penalty Total, which is in turn needed to calculate the Raw Quality Score and the Calibrated Quality Score.

MA Evaluation scorecard (after core MQM 2.0)

Client	Project Name		Translator		Evaluator	
	Master's thesis				Sophie Grossalber	
Grossalber	Source Language	Target Language		Total Word Count	Evaluation Word Count (EWC)	Content Type
	EN-us	DE-at			299	NATO website text
Score Calculation Parameters	Maximum Score Value (MSV)	Passing Threshold Raw QS	Passing Threshold Calibrated QS		Reference Word Count (RWC)	Acceptable Penalty Points/RWC
	100	99	90		1000	10
Evaluation Results	Quality Rating (QR)	Raw Quality Score (RQS)	Absolute Penalty Total (APT)	Per-Word-Penalty Total (PWPT)	Normed Penalty Total (NPT)	Calibrated Quality Score (CQS)
	PASS	100	0	0	0	100

General Comment

Commentary...

Severity Penalty Multiplier	0	1	5	25				
Error Type	Neutral Errors Count	Minor Errors Count	Major Errors Count	Critical Errors Count	Error Type Penalty Total (ETPT)	Weighted ET Penalty Points	Normed ET Penalty Points	Error Type Weight
Terminology	0	0	0	0	0	0	0	2
Accuracy	0	0	0	0	0	0	0	1
Linguistic conventions	0	0	0	0	0	0	0	1
Style	0	0	0	0	0	0	0	1
Locale conventions	0	0	0	0	0	0	0	1
Audience apropriateness	0	0	0	0	0	0	0	2
Grand Total	**0**	**0**	**0**	**0**	**0**	**0**	**0**	

Figure 1: Base scorecard for the evaluation of translations used for this thesis, modelled after core MQM 2.0.

As the Evaluation Word Count is usually based on the source text, the evaluation word count is thus set to 299, which is the total word count for the piece of the NATO website that was chosen for translation and evaluation. The actual word counts for the DeepL and ChatGPT translations are 316 and 286, respectively. The human translation runs to 310 words. The Reference Word Count was set at 1000 words, the industry default. The Passing Threshold for the Raw Quality Score was set to 99, and the Passing Threshold for the Calibrated Quality Score was set to 90, which leaves a total of 10 Acceptable Penalty Points per Reference Word Count. The Defined Passing Interval needed for the calculation of the Scaling Factor and the Calibrated Quality Score was then also set to 10. A translation that achieved, i. e., 99.3 points for the Raw Quality Score and 93.3 points for the Calibrated Quality Score, would pass the evaluation.

Due to the nature of the source text and the intended function of both source and target texts (inform readers who are most likely familiar with military and NATO terminology), the decision was made to weigh errors falling into the **Terminology** or **Audience Appropriateness** categories more heavily than errors in other categories. The Error Type Weight for those two categories was set to 2, whereas the Error Type Weight for the other categories was set to the default of 1. The Severity Penalty Multiplier was set to the MQM recommended default of 0 for neutral errors, 1 for minor errors, 5 for major errors and 25 for critical errors. One critical error would result in an immediate failure of the evaluation for that text. As the evaluation of one's own translation proves difficult and the author of the thesis did not have access to the terminology resources of the AAF's Language Institute, the human translation was done by a translator employed by the AAF. Afterwards, an expert interview was conducted with this same translator to answer any questions regarding translation strategy and terminology questions.

For the evaluation of the target texts, the *Militärwörterbuch – Englisch* (Landesverteidigungsakademie des österreichischen Bundesheeres 2014) and the glossary on the Army in Europe and Africa Publications' website (Army in Europe and Africa Publications 2024) were used for cross-referencing correct or incorrect terminology.

2.2.2. Interview setup

Two different interview setups were used for this thesis. First, an interview was conducted with the AAF translator with questions regarding their translation experience, their prior experience with machine translation and some questions on terminology and the used choices. The interview was conducted in person and recorded with a digital voice recorder. The transcript was then made using the speech-to-text feature in MS Word and manual edits of any speech-to-text errors. The interview's audio recording is 36 mins and 30 secs long, the transcript runs up to nine pages. Relevant passages, which might be cited in part in any of the thesis' Chapters can be found as a whole in the appendix and are cited with "T".

The second interview type was a qualitative interview divided into three sections and conducted face-to-face with one person at a time, five persons in total. Interviews were planned to, ideally, last about 30 minutes and divided into three parts. In the first part, questions about the interviewee's prior experience with translation were asked, i.e. whether they had experience with translations due to their position within the Austrian Armed Forces and the Ministry of Defence, and whether they had ever used DeepL, ChatGPT or another

LLM or machine translation model to translate a text and their conclusions to those experiences if they had any.

Part two of the interview constituted the main part of the experiment with the participants being asked to rank the translations they were presented with based on what they thought was more likely machine translated or human translated. Participants at this point were not yet told which text was translated from English into German with ChatGPT, DeepL Translator or by the expert from the AAF. After their ranking, they were asked to justify their choices. The goal here was to see how their previous experience with machine translation – or the lack thereof – would influence their perception of the texts and whether they would be able to catch some of the errors made by the machine translation. As explained in Chapter 1, this thesis is a pilot study into the use of and perception of machine translation within the Austrian Armed Forces, with the possibility of serving as a basis for the development of training for both translators and AAF personnel. The reasons, i. e. possible security concerns around the use of LLMs and machine translation, have been explained in the previous Chapter.

In the last interview part, participants were told which of the texts they ranked were translated with ChatGPT, DeepL Translator or by the human expert and asked questions on if and how this knowledge would change their ranking and their perception of the texts.

The expert interview was conducted in English. The interviews with the ranking component were conducted in German to facilitate easier discussion of the translated texts and to avoid the creation of an entry barrier for those below a certain level of English. As with the expert interview, transcripts of the ranking interviews can also be found in the appendix – as well as the translated texts in full. Both the expert interview as well as the experiment interviews were then transcribed using the speech-to-text transcription function built-in in MS Word to save time, and then edited afterwards to correct any errors made by the speech-to-text module. Quotes of the experiment interviews used in the thesis were translated with DeepL and then post-edited where necessary. The transcript of the interviews with the experiment yielded more errors than the expert interview, presumably because the expert interview was conducted in English and the interviews with the experiment component were conducted in German. In addition to the languages that the interviews were conducted in, MS Word's speech-to-text function only includes the option for the German locale for German and not the Austrian one. The function then appeared to have trouble understanding the Austrian accents

and proceeded to misspell ChatGPT and DeepL. In addition to misunderstanding accents, it also misunderstood instances of “Bundesheer” and wrote “Bundeswehr” instead, which led me to believe that it had been primarily trained with texts or speeches containing “Bundeswehr”. There had been a total of 25 instances where the text-to-speech function misheard ChatGPT, and 23 where it misheard DeepL over the five interview transcripts that contained the experiment component. Some of the instances for “ChatGPT” included words like:

- jetgip
- chip bt
- Jet GPT
- Chair GPT
- Chad GPT

The list for “DeepL” on the other hand, included words like:

- Deep Error
- Deep R
- Deep Pearl
- Deep eye
- Dipl.

However, given that the language of the speech files that the speech-to-text function was given to transcribe, was German, “Dipl.” could at least be explained with being part of the abbreviation for “Diplom-Ingenieur/Diplom-Ingenieurin”, which is “Dipl.-Ing.”. (Bundesministerium für Bildung, Wissenschaft und Forschung 2024) The other instances for both DeepL and ChatGPT are likely to be traced back to the accents of the speakers, which were predominantly Austrian, and also the audio quality of the recordings.

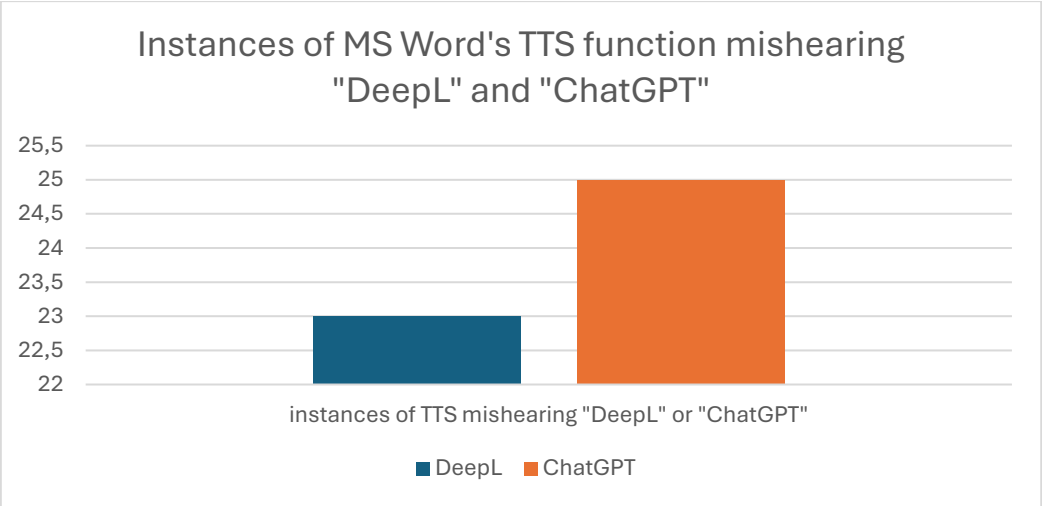

Figure 2: Instances of MS Word's TTS function mishearing "DeepL" and "ChatGPT"

2.3. Text Analysis

In this section, I will briefly analyse the source text and compare the profiles of the source and target texts. Analysis and profile models are based on Christiane Nord's models as she had laid them out in her book *Text Analysis in Translation: Theory, Methodology, and Didactic Application of a Model for Translation-Oriented Text Analysis*. (2006)

2.3.1. Source and Target Text Profile Comparison

As there is no official, external translation commission apart from the translation commission I gave the AAF translator for production of the human translated target text, I will be analysing and comparing the Source and Target Text Profiles solely with NATO's website. Due to this, the motivation behind the translation of the Source Text will differ from an official translation commission as the motivation behind the thesis is to investigate how well machine translation could translate a text in the military domain into the Austrian locale of German and how those translations might be perceived by the Austrian target audience.

2.3.1.1. Source Text Profile

Based on Nord's list of information the translation commission – or sometimes brief –, we can infer the following pieces of information about the source text. As there is no external translation commission from an official client, the source text and the motivation behind this thesis will act as guidelines for this list.

The **motive** for as to why the source text was written is NATO's desire to inform visitors to their site and readers on why and how NATO supports Ukraine in its defence

against the Russian invasion, as well as combating malicious disinformation that has been propagated by Russia and other actors. The motive for a **translation** into German could then be to make this information accessible to German native speakers without the proper language skills to read the original site in English or the already existing translations in French, Ukrainian and Russian. The motive for a translation from my point of view has already been touched upon in subchapter 2.3.1.

From the motive we can then derive the intended **text function**(s), which in this case are **informing** the readership and relaying information to them in regard to NATO's reaction to Russia's war of aggression against Ukraine. The text also functions as one piece of the fight against Russian disinformation.

The **sender** is clearly NATO itself, and the **addressees** are readers of various cultural backgrounds and origins, with the added hindrance of the text's languages only being English, French, Ukrainian Russian. There is no discernible distinction made between age groups of possible readers or their gender identities as the text is intended for a large audience.

The **time** of reception depends entirely on when the reader comes across the website and is not linked to a specific time zone. The **place** of reception is equally broadly defined as it hinges on the reader's location.

As for **medium**, the text is available in written form and digitally, although the reader could also print out the page if so desired.

2.3.1.2. Target Text Profile

As with the Source Text Profile, the Target Text's **motive** is the desire to inform visitors to NATO's website on the war of Ukraine and steps taken by NATO and its allies to support Ukraine in its war. Translating the text into German would grant accessibility to readers with no or little knowledge of English, French, Ukrainian or Russian.

The intended **text functions** are the same as for the Source Text: to inform the readership of steps taken by NATO and its allies in its support of Ukraine in the war. Sender, addressees, time and place of reception and medium are also the same for the Source Text.

2.3.2. Source Text Analysis

As mentioned, the source text is a section called "The Comprehensive Assistance Package (CAP)" on NATO's own website. (NATO 2024) The text is originally available in English

with British spelling and the site offers translations into French, Russian and Ukrainian. There is no official German translation of the text available on NATO's website.

The **subject matter** is NATO's relationship with Ukraine and the organisations response to Russia's war of aggression against Ukraine. As NATO is an international organisation with different countries as members, the content on the site in question isn't culture-bound in the sense that it is bound to the cultural conventions of the source language as English here acts as lingua franca, as does French.

The **content** is not solely background information on the situation in Ukraine and a history of NATO's relationship with Ukraine, but also constitutes information on the steps taken by NATO to support Ukraine and NATO's general response to the war. The content's connotation is a clear signal to both NATO Allies, partners and Ukraine, as well as Russia and non-Allies and non-partner countries NATO's commitment to long-term support of Ukraine and the solidarity shared among Allies.

The text does **presuppose** a certain level of familiarity from the reader with military terminology, particularly in English – or in the target language. As the target audience is expected to be familiar with – or have a background in – the military and the organisational structure of NATO, certain terms are also used in their official abbreviated form, like "C4", which does not refer to the explosive device in this case, but a military concept of command structures. (NATO 2024)

In terms of **text composition**, the site is clearly structured and divided into subsections with headings and also **non-verbal elements** like text boxes that are set apart from the main text with a different background colour and give an overview of the most important points. The site also moves chronologically through the text from top to bottom and according to which information is more important to readers, with NATO-Ukraine relations at the top in the grey box as a bulleted list, followed by a section on practical support to Ukraine and closing with information on Ukraine's aspirations for NATO membership. There are also some hyperlinks in later sections which lead to other NATO sites with longer explanations where applicable. The section used for the experiment also uses sub-headings and a bulleted list for text composition and structuring.

As for **Lexis**, the text employs a fairly large amount of subject-specific terminology, which would necessitate the use of a translator or editor (ideally both) with the corresponding level of expertise in the military domain to solve translation problems with terminology. For

example, with “Command, Control, Communications and Computers (C4)” (par. 4) as mentioned in the section on presuppositions, “command” and “control” are usually translated into German as “Führung” and, thus, the term “Führung” takes on the task of communicating the concepts behind two separate terms in English. For register, the text is aimed at readers with a level of English equivalent to at least C1 within the CEFR (~ Level 2+ or higher within STANAG 6001), it is written in formal English and uses British English spelling.

In terms of **sentence structure**, the text exhibits mostly long sentences, with even the bullet points being fairly long and extensive, where appropriate a lot of commas were used. Some defining clauses were used, but there are also on-defining ones. Ther is no use of ellipsis and parenthesis were only used to add abbreviations of organisation names or concepts where those were first mentioned, i. e. “United Nations (UN)”. Abbreviations were used consistently throughout the text after the first instance of appearance. Use of gender neutral or gendered terms is not consistent throughout the text, see “servicemen and women and civilian personnel”. (NATO 2024: 5)

Suprasegmental features are not applicable to this text as it is only available in written form. There is no emphasis on particular words or phrases through choice of writing style.

3. Translation Evaluation

I have laid out the reasons and methodology for the translation evaluation in Chapter 2 and will now discuss the findings. The hope is that my findings on translation errors and the quality of the machine translations by ChatGPT and DeepL, as well as the evaluation of the human translation, will match up with the answers given by participants during the interviews. As mentioned in Chapter 2, errors in the **Terminology** and **Audience Appropriateness** categories were weighted at 2, while other error types were weighted at the default of 1. This was done simply due to the importance of correct terminology and audience appropriateness in a text within the military domain. While accuracy and linguistic conventions should be held in equally high importance, minor grammar errors can be overlooked by native speakers and the target audience. Terminology errors are trickier, not only to classify as there is no official dictionary shared between Germany, Austria and Switzerland of military terms, but also because they can severely hinder understanding. That said, none of the translations feature critical terminology or audience appropriateness errors which would mean an immediate failure of the evaluation. And the machine translation and LLM translation were both used without post-editing. Their raw output was evaluated, with post-editing some errors might have been eliminated altogether. However, neither the translation done by DeepL nor the one done by ChatGPT passed my evaluation.

3.1. Human Translation and Expert Interview

The human translation with its word count of 310 words sits between DeepL and ChatGPT's translations in length. The Passing Threshold for the Raw QS was set to 99, with a Maximum Value Score of 100 that leaves a difference of 1. The number of Acceptable Penalty Points per Reference Word Count was then set to 10, in accordance with the Passing Threshold for the Raw Quality Score and the Passing Threshold for the Calibrated Quality Score is thus 90 points. Although this translation was used as something of a "gold standard" for the machine translations, I did find one neutral and one minor style error, which brought the final Raw Quality Score to 99.665% and the Calibrated Quality Score to 96.65%. This constitutes an overall PASS grade for the human translation. The human translator "T" possesses the necessary background for a specialised translation within the military domain and the resources for terminology research specific to the Austrian Armed Forces. They were not given special instructions, i. e. not to use machine translation, apart from the target audience being Austrian and of military background.

MA Evaluation scorecard (after core MQM 2.0)

Client	Project Name		Translator		Evaluator	
	Master's thesis		Human		Sophie Grossalber	
Grossalber	Source Language	Target Language		Total Word Count	Evaluation Word Count (EWC)	Content Type
	EN-us	DE-at			299	NATO website text
Score Calculation Parameters	Maximum Score Value (MSV)	Passing Threshold Raw QS	Passing Threshold Calibrated QS		Reference Word Count (RWC)	Acceptable Penalty Points/RWC
	100	99	90		1000	10
Evaluation Results	Quality Rating (QR)	Raw Quality Score (RQS)	Absolute Penalty Total (APT)	Per-Word-Penalty Total (PWPT)	Normed Penalty Total (NPT)	Calibrated Quality Score (CQS)
	PASS	99,66555184	1	0,003344482	3,344481605	96,65551839

General Comment

inconsistent use of genderneutral or gendered speech - neutral inconsistent style error; ability and capability both translated as Fähigkeit - although correct, reads awkwardly - minor awkward style error

Severity Penalty Multiplier	0	1	5	25				
Error Type	Neutral Errors Count	Minor Errors Count	Major Errors Count	Critical Errors Count	Error Type Penalty Total (ETPT)	Weighted ET Penalty Points	Normed ET Penalty Points	Error Type Weight
Terminology	0	0	0	0	0	0	0	2
Accuracy	0	0	0	0	0	0	0	1
Linguistic conventions	0	0	0	0	0	0	0	1
Style	1	1	0	0	1	1	0,003344482	1
Locale conventions	0	0	0	0	0	0	0	1
Audience apropriateness	0	0	0	0	0	0	0	2
Grand Total	**1**	**1**	**0**	**0**	**1**	**1**	**0,003344482**	

Figure 3: Scorecard of evaluated human translation

The first thing that stood out with the human translation was the high similarity in the first sentence between the human and DeepL's machine translation which was not mentioned in the expert interview, and thus suggests that DeepL might have been used in the human translation with heavy post-editing for other parts of the text.

During the interview, T was asked why they translated certain terms and not others. "Best practices" (Human Translator "T" 2024: 1) was not translated, while DeepL and ChatGPT did translate it into German, as there is no true German equivalent, and the English term has found its way into everyday use in German. When asked about why T translated some terms and not others, they said: "So C2, C3, C4, C4 ISR - these are all terms, nobody translates this stuff, but I don't even think anybody could agree on a translation. It's just, it's just thrown in. Because you asked me to translate this, I, of course, then started translating it. And I ended up finding German home pages of the German Bundeswehr where it said, 'Das könnte man so sagen. Ungefähr'." (Expert interview 2024: 33) They clearly cited current locale conventions and translation practices as the reason why some terms would not be translated. The added assumption that nobody could agree on a translation was one I also reached while researching the source text and finding no official German translation of the NATO website. Translator T was also not given any specific instructions to leave certain

terms untranslated, but their mention of the German Bundeswehr having reached an acceptable translation suggests that there have been efforts to develop the terminology for German as well, likely because not every German native speaker also speaks English well enough to understand what was meant. Particularly with a specialised text such as this one, whose function is clearly communicative and informative, for those in the target audience who do not understand English well enough, the existence of somewhat agreed upon terminology in German is an appropriate strategy to mitigate possible issues with comprehension and misunderstandings.

Although T noted that certain concepts and terms are simply not translated anymore, and the original English language term is used instead, some terms were translated according to AAF conventions ("Kampfmittelbeseitigung", "Abwehr behelfsmäßiger Spreng- und Brandvorrichtungen"). However, they kept the English abbreviations. "Command and control" was translated as "Führung" and, when asked, translator T said:

> *'Führung' is a wonderful term in German because it covers everything. In English you have to differentiate between leadership and command and control. Leadership is always something – the way we define it as how you interact with people. […] And then, of course, there's command and control. Now most German speakers have no idea what controlling means. They translate controlling as 'kontrollieren' which it is not, it's 'nachsteuern'. […] The question is, when an order comes in, was implemented and then you have to check how has this order been implemented, how has the situation changed. That's of course the C2 process. (Expert interview 2024: 22)*

Translator T also saw precisely these translation problems as potential challenges for machine translation models. As, according to them, it is assumed that machine translation models are very good at technical and specialised translations but there are so many differences in American, British, German and Austrian military terminologies that a translation of a certain requires the translator to be able to perform terminology research and to decipher the context correctly.

The human translation also utilizes the original, English language abbreviation for the terms that have been translated into German, in line with linguistic conventions and as is appropriate for the target audience.

The text does not consistently make use of gender-equitable language in German ("Patienten", "Soldatinnen und Soldaten") as would be required in the regulation GZ S90100/6-S I/2018. (Mathes & Gruber 2019: 43) This regulation allows the use of a disclaimer that men and women are included in male terms in pieces of writing or the use of

female and male terms. The disclaimer afterwards requires the consistent use of the generic masculine in a text. The machine translations do not follow this regulation either. Where the machine translation models' inconsistent use of gender strategies in the translation could be excused by arguing that the models do not have enough training data to warrant such choices being made, the human translator does not have an excuse to that extent. This also did not come up in the expert interview conducted with translator T. As translator T themselves said in the interview: "Words are suddenly very much out of fashion because it no longer reflects the world we want to create linguistically. So, you have to be very flexible, and you have to be constantly aware of changes." (Expert interview 2024: 21) Which would also require the translator being aware of any guidelines or rules toward gender-neutral language use while translating for a specialised audience like the Austrian Armed Forces.

> *The one theory about machine translation is that the machine translation is really, really good at technical texts or at technical language, in the sense that it's a specialised set of terminology. No, it's not. It's getting better there as well but what people who underestimate, especially about military text or texts related to the military sphere, is how much politics plays a role there. (Expert interview 2024: 19)*

According to translator T, it is precisely these politics that provide challenges for machine translation when translating military texts as the MT models cannot grasp the nuances required for a qualitative adequate and useful translation. This problem could, of course, be easily mitigated with post-editing after the machine translation, which would ideally be done by a translator or post-editor trained in, and with experience in, the military domain and the translation of such texts. Whether translator T managed to produce a translation that was adequate for the required purpose and the study participants acting as the target audience, will be discussed in Chapter 3.2, the presentation of interview results.

3.2. Evaluation Results of DeepL's translation

DeepL's translation is 316 words long and thus the longest out of all three translations. The machine translation output was not post-edited prior to evaluation or the interviews with the study participants. The Passing Threshold for the Raw QS was, again, set to 99, with a Maximum Value Score of 100 with a difference of 1. The number of Acceptable Penalty Points per Reference Word Count was then set to 10, in accordance with the Passing Threshold for the Raw Quality Score and the Passing Threshold for the Calibrated Quality Score is thus 90 points. As has been already mentioned, DeepL's translation failed the evaluation with a final Raw Quality Score of 98.662% and a Calibrated Quality Score of

86.622%. The Absolute Penalty Total (APT) amounts to 4, with the Per-Word-Penalty Total being 0.013377926. The Normed Penalty Total is 13.377924642.

MA Evaluation scorecard (after core MQM 2.0)

Client	Project Name		Translator		Evaluator	
	Master's thesis		DeepL		Sophie Grossalber	
Grossalber	Source Language	Target Language		Total Word Count	Evaluation Word Count (EWC)	Content Type
	EN-us	DE-at			299	NATO website text
Score Calculation Parameters	Maximum Score Value (MSV)	Passing Threshold Raw QS	Passing Threshold Calibrated QS		Reference Word Count (RWC)	Acceptable Penalty Points/RWC
	100	99	90		1000	10
Evaluation Results	Quality Rating (QR)	Raw Quality Score (RQS)	Absolute Penalty Total (APT)	Per-Word-Penalty Total (PWPT)	Normed Penalty Total (NPT)	Calibrated Quality Score (CQS)
	FAIL	98,66220736	4	0,013377926	13,37792642	86,62207358

General Comment

Overall technically adequate translation. Although C4 could be translated, the English term is also used in German - the translation still counts as a minor terminology error as it does not affect comprehension too much for the target audience that would be familiar with the concepts. BUT "control" cannot be translated as "Kontrolle" in this context. Command and control are usually translated as "Führung" as the concept relates to leadership and reviewing the command processes. The misspelling of "CAP" as "GAP" has been categorised as a minor spelling error as it would potentially hinder comprehension but it is an easily fixable error during post-editing. Although the DeepL text does a lot better terminology-wise than ChatGPT and also has a lower NPT score, it is still considered a fail from the score alone. The ChatGPT text also reads more idiomatic in some cases. A combination of the two texts with a thorough post-edit with a focus on terminology would be ideal.

Severity Penalty Multiplier	0	1	5	25				
Error Type	Neutral Errors Count	Minor Errors Count	Major Errors Count	Critical Errors Count	Error Type Penalty Total (ETPT)	Weighted ET Penalty Points	Normed ET Penalty Points	Error Type Weight
Terminology	3	1	0	0	2	2	0,006688963	2
Accuracy	1	1	0	0	1	1	0,003344482	1
Linguistic conventions	4	1	0	0	1	1	0,003344482	1
Style	3	0	0	0	0	0	0	1
Locale conventions	0	0	0	0	0	0	0	1
Audience appropriateness	0	0	0	0	0	0	0	2
Grand Total	**11**	**3**	**0**	**0**	**4**	**4**	**0,023411371**	

Figure 4: Scorecard of evaluation of the DeepL translation

Within the DeepL translation there were a grand total of 11 Neutral Errors, those are errors that are small enough not to impact the comprehension of the target text and thus do not count towards the Penalty Totals. There were also 3 Minor Errors, with 1 in Terminology, 1 in Accuracy and 1 in Linguistic Conventions. Minor Errors are multiplied with a Penalty Multiplier of 1 and, while they do impact the comprehension of the text somewhat and thus count towards the Penalty Totals, their impact is small enough that native speakers might catch the errors but the meaning is not influenced too much. There were no Major or Critical Translation Errors.

Terminology Errors were weighted with 2 while other categories were weighted with 1, due to the fact that precise communication is important during international operations but also with the Austrian Armed Forces' own personnel in their own language. The agreement to use standardised terminology for shared concepts is a necessity and thus also requires enforcing of the use of said terminology to avoid misunderstandings – which could prove to be fatal to personnel and for the outcome of a mission. Additionally, the text functions of both

the source and target texts are communicative and informative, and thus also required careful consideration of phrasing and terminology. Due to the fact that DeepL does not offer the Austrian locale for German and was likely not trained on German texts from the Austrian Armed Forces, terminology errors are to be expected and require a post-editor with knowledge of the locale and the domain, as well as the resources to look up terminology as needed. Those resources were also used for the evaluation.

Severity Penalty Multiplier	0	1	5	25
Error Type	Neutral Errors Count	Minor Errors Count	Major Errors Count	Critical Errors Count
Terminology	3	1	0	0
Inconsistent w/ term. resource	0	0	0	0
inconsistent use of terminology	0	0	0	0
wrong term	3	1	0	0
Accuracy	1	1	0	0
mistranslation	0	1	0	0
overtranslation	0	0	0	0
undertranslation	0	0	0	0
addition	0	0	0	0
omission	0	0	0	0
do not translate	1	0	0	0
untranslated	0	0	0	0
Linguistic conventions	4	1	0	0
grammar	4	0	0	0
punctuation	0	0	0	0
spelling	0	1	0	0
unintelligible	0	0	0	0
textual conventions	0	0	0	0
Style	3	0	0	0
organization style	0	0	0	0
third-party style	0	0	0	0
inconsistent with ext. Ref.	0	0	0	0
language register	1	0	0	0
awkward style	0	0	0	0
unidiomatic style	0	0	0	0
inconsistent style	2	0	0	0
Locale conventions	0	0	0	0
Audience apropriateness	0	0	0	0
Grand Total	**11**	**3**	**0**	**0**

Figure 5: Detailed look at Error Type Totals for DeepL

The term C4 (Command, Control, Communication, Computers) was translated as “Kommando, Kontrolle, Kommunikation und Computer”. As mentioned in subchapter 3.1.1., a translation of the term is not strictly necessary as the concept and the English term are also used by the Austrian Armed Forces. However, if a translation is required or demanded, translating “Command, Control” literally as “Kommando, Kontrolle” can lead to confusion within the target audience. These two terms are usually condensed into one and translated as “Führung” as “Command, Control” relate to leadership, leadership processes and the control

processes of the leadership processes. “Kommando” can be misunderstood as an organisational unit within the Austrian Armed Forces, i. e. “Militärkommando Niederösterreich”, or a person being in command of personnel in the field. And “Kontrolle” could perhaps be mistaken for taking control of a point or an area or controlling a person. Neither translations are effective for the intended purpose of communicating the concept behind the English term. However, including the abbreviation of C4 in brackets behind the term serves to mitigate any risks of misunderstanding for people familiar with the concept itself, the English term and the abbreviation commonly used for it. “Kontrolle” has been classified as a Minor Mistranslation Error, due to the possibility of misunderstandings arising from the use in this particular domain and context.

The misspelling of “CAP” as “GAP” has been classified as a minor spelling error as it would hinder comprehension and foster confusion with the target audience as the abbreviation for the Comprehensive Assistance Package has been previously introduced as “CAP”. However, these misunderstandings could be resolved quickly, and it is to be assumed that an official publication would not use raw machine translation output without post-editing prior to publication.

There were two Neutral Style Errors to do with inconsistent style. One was the inconsistent use of gendered or gender-neutral language (“Patienten”, “Soldatinnen und Soldaten”, “Personal” in paragraph 5). But, as mentioned briefly in subchapter 3.1.1., it is likely that DeepL possesses more training data outside of the military domain where texts used “Patienten” instead of “Patientinnen und Patienten” and “Personal” for the translation from English. Additionally, the machine translation model is unable to research and read certain regulations within an organisation as a human translator would likely do. The other instance of a neutral inconsistent style error was a case of a missing bullet in the bulleted list. (DeepL Translator 2024: 5)

“The destruction of small arms and light weapons (SALW), conventional ammunition, and anti-personnel landmines” (NATO 2024: 7) was translated by DeepL as “Vernichtung von Kleinwaffen und leichten Waffen (SALW), konventioneller Munition und Antipersonenminen” (DeepL Translator 2024: 7). Initially, “leichte Waffen” was classified as

a Neutral Terminology Error, however, research showed that both “leichte Waffen” and “Leichtwaffen” can be used and the concept behind the terms remains unchanged.[3]

The same is true for “Cyberverteidigung” in the same paragraph. The original term is “cyber defence” (NATO 2024: 7) and can be translated as either “Cyberverteidigung” or “Cyberabwehr”. The use of “Cyberverteidigung” in this context does not constitute an error per se, the word “Verteidigung” when translating “defence” simply reads unidiomatic. The AAF predominantly use “Cyberabwehr”.

On sentence level, there were more than one sentence that shared similarities with the human translation, such as the first sentence in paragraph 1:

> “Auf dem NATO-Gipfel 2016 in Warschau wurden die Maßnahmen des Bündnisses zur Unterstützung der Ukraine Teil des Umfassenden Beistandspakets (Comprehensive Assistance Package, CAP), mit dem die Fähigkeit der Ukraine unterstützt werden soll, für ihre eigene Sicherheit zu sorgen und weitreichende Reformen auf der Grundlage von NATO-Standards, euro-atlantischen Grundsätzen und bewährten Praktiken durchzuführen.” (DeepL Translator 2024: 1)

But also in paragraph 2, where the sentences read similarly to the human translation except for some grammatical and terminological differences. The grammatical differences can be traced back to DeepL’s misspelling of “CAP” as “GAP” and subsequently misgendering the Comprehensive Assistance Package, which would be neutral in German and require the article “das”, as feminine with the article “die”.

Overall, DeepL’s translation is an adequate translation, but it would still require heavy post-editing and a terminology check by an expert within the domain and the locale, who is aware of the locale conventions and terminology usage. From the score alone, the translation is still considered a fail. While the difference in RQS from the human translation and DeepL’s translation is only 1%, giving Terminology Errors a weight of 2 instead of 1, still pushes its score below the Passing Threshold.

3.3. Evaluation Results of ChatGPT’s translation

With 286 words, ChatGPT’s translation is the shortest out of the three texts and also shorter than the source text with 299 words. As with the DeepL translation, ChatGPT’s text was not post-edited before evaluation or the interviews with the study participants. The Passing

[3] See, for example, https://www.auswaertiges-amt.de/de/aussenpolitik/sicherheitspolitik/abruestung-ruestungskontrolle/uebersicht-konvalles-node/kleinleichtwaffen-node [last accessed October 20th, 2024]

Threshold for the Raw QS was also set to 99, with a Maximum Value Score of 100 with a difference of 1. The number of Acceptable Penalty Points per Reference Word Count was then set to 10, in accordance with the Passing Threshold for the Raw Quality Score and the Passing Threshold for the Calibrated Quality Score is thus 90 points. ChatGPT's Raw Quality Score was ultimately 94.98%, and the Calibrated Quality Score 49.832%. The Absolute Penalty Total was 15, the Per-Word-Penalty Total 0.05016. The Normed Penalty Total was 50.1672. With those scores for Raw Quality and Calibrated, ChatGPT's translation still fails the evaluation.

MA Evaluation scorecard (after core MQM 2.0)

Client	Project Name		Translator		Evaluator	
	Master's thesis		ChatGPT		Sophie Grossalber	
Grossalber	Source Language	Target Language		Total Word Count	Evaluation Word Count (EWC)	Content Type
	EN-us	DE-at			299	NATO website text
Score Calculation Parameters	Maximum Score Value (MSV)	Passing Threshold Raw QS	Passing Threshold Calibrated QS		Reference Word Count (RWC)	Acceptable Penalty Points/RWC
	100	99	90		1000	10
Evaluation Results	Quality Rating (QR)	Raw Quality Score (RQS)	Absolute Penalty Total (APT)	Per-Word-Penalty Total (PWPT)	Normed Penalty Total (NPT)	Calibrated Quality Score (CQS)
	FAIL	94,98327759	15	0,050167224	50,16722408	49,83277592

General Comment

"Alliierte" for "allies" in the context of Russia's war against Ukraine ahs been vategorised as a minor wrong term error as "Alliierte" is usually only used in context of World War 2, and the use of this term in the target context would evoke wrong associations with the target audience. "Kontrolle" has been categorised as a minor wrong term error as well as it does not affect comprehension with the target audience too much, but "control" from C4 is not translated as "Kontrolle" but as "Führung" together with "Command" (see scorecard DeepL).An acceptable translation would be "Führung, Information, Kommunikation, Computersysteme", but usually the English term is used. "Dienstmänner und -frauen" was categorised as a major wrong term error as it is a literal translation from the source text and "Dienstmann" means a vassal or somebody who carries luggage and not, what is actually meant in this context, soldiers. Although, terminology-wise, the translation from ChatGPT is a fail and fails with a bigger NPT score than the DeepL translation, it is at times more idiomatic than the DeepL text.

Severity Penalty Multiplier	0	1	5	25				
Error Type	Neutral Errors Count	Minor Errors Count	Major Errors Count	Critical Errors Count	Error Type Penalty Total (ETPT)	Weighted ET Penalty Points	Normed ET Penalty Points	Error Type Weight
Terminology	1	2	1	0	14	14	0,046822742	2
Accuracy	0	0	0	0	0	0	0	1
Linguistic conventions	1	1	0	0	1	1	0,003344482	1
Style	4	0	0	0	0	0	0	1
Locale conventions	0	0	0	0	0	0	0	1
Audience apropriateness	0	0	0	0	0	0	0	2
Grand Total	**6**	**3**	**1**	**0**	**15**	**15**	**0,050167224**	

Figure 6: Scorecard of the evaluation of ChatGPT's translation

There were a total of six Neutral Errors, with one in terminology, two in Linguistic Conventions and three in Style. The Minor Errors Count amounts to three, with two in Terminology and one in Linguistic Conventions. Then, there is also one Major Error in Terminology. Due to the weighting of Terminology with 2 instead of 1, the Error Type Penalty Total for that category is 14, while the Error Type Penalty Total for Linguistic conventions is only 1.

Severity Penalty Multiplier	0	1	5	25
Error Type	Neutral Errors Count	Minor Errors Count	Major Errors Count	Critical Errors Count
Terminology	1	2	1	0
Inconsistent w/ term. resource	0	0	0	0
inconsistent use of terminology	0	0	0	0
wrong term	1	2	1	0
Accuracy	0	0	0	0
Linguistic conventions	1	1	0	0
grammar	1	1	0	0
punctuation	0	0	0	0
spelling	0	0	0	0
unintelligible	0	0	0	0
textual conventions	0	0	0	0
Style	4	0	0	0
organization style	0	0	0	0
third-party style	0	0	0	0
inconsistent with ext. Ref.	1	0	0	0
language register	0	0	0	0
awkward style	1	0	0	0
unidiomatic style	0	0	0	0
inconsistent style	2	0	0	0
Locale conventions	0	0	0	0
Audience apropriateness	0	0	0	0
Grand Total	**6**	**3**	**1**	**0**

Figure 7: Detailed look at Error Type Totals for ChatGPT

The first paragraph of ChatGPT's translation already includes an error as it translated "Alliance" as "Allianz". While technically correct, NATO is not spoken of in German as an alliance but as "Bündnis". This then was counted as a Neutral Style Error in the subcategory "inconsistent with external reference" as seen in Figure 5. While the human translator kept "best practices" in paragraph 1 as "best practices", both DeepL and ChatGPT translated it as "bewährte Praktiken". Again, technically a correct translation, but as "best practices" has found its way into the everyday use of German as the English term, it was counted as a Neutral Style Error in the category "awkward style".

In paragraph 2, ChatGPT translated "NATO and its allies" as "NATO und ihre Alliierten", which was counted as a minor wrong term error as "Alliierte" is usually associated with the Allied Forces in World War 2 in Germany and Austria. This error, if not caught during post-editing, would most definitely lead to confusion among the target audience as it would evoke the wrong connotation and associations with the word. The correct translation would be "NATO und ihre Verbündeten", as NATO-partner nations are usually called.

As with DeepL, ChatGPT translated “Command, Control, Communications, Computers” (NATO 2024: 5) as “Führung, Kontrolle, Kommunikation und Computer” (ChatGPT-4 2024: 5). While DeepL made the mistake of translating “command” as “Kommando”, ChatGPT did correctly translate it as “Führung”, but left out “control” and instead translated that wrongly as “Kontrolle”. As explained in subchapter 3.1.2., “command, control” are usually summarised in German as the term “Führung”. This translation error has been classified as a minor wrong term error.

The only Major Wrong Term Error in ChatGPT’s translation, which significantly impacts meaning and the final Quality Scores, is “Dienstmänner und -frauen” found in paragraph 6. The original in the Source Text is “service men and women”, which generally translates to “Soldatinnen und Soldaten” in German. “Dienstmann” could by either a vassal or a porter and was most often used in German until the 1950s, according to the *Digitales Wörterbuch der deutschen Sprache* (DWDS – Digitales Wörterbuch der deutschen Sprache 2016).

One Neutral Grammar Error can also be found in paragraph 6, as the article “die” in the sentence “Medizinische Rehabilitation, die darauf abzielt, die Ukraine bei der Verbesserung ihres Systems der medizinischen Rehabilitation zu unterstützen, um sicherzustellen, dass langfristig nachhaltige Dienstleistungen für Patienten, einschließlich aktiver und entlassener ukrainischer Dienstmänner und -frauen sowie ziviler Mitarbeiter des Verteidigungs- und Sicherheitssektors, erbracht werden;” relates back to „medizinische Rehabilitation” when it should actually reference back to „Treuhandfondsprojekte”.

Same as with the human translation and DeepL, ChatGPT does not consistently use gendered or gender-neutral language. However, this could have been reflected in the prompt, as opposed to DeepL’s translation, where that option does not exist. The focus of the experiment was, however, on terminology and fluency of the text. During post-editing, this would have had to be addressed, of course. Apart from “Patienten” and “Dienstmänner und -frauen” in paragraph 6, other parts of the translation use the generic masculine, like “Mitarbeiter” in paragraph6 or “Zivilisten” in paragraph 7.

Regarding terminology, ChatGPT’s translation has made more, and more severe, errors than DeepL’s translation, which is also reflected in the Absolute Penalty Total and the Normed Penalty Total as well as the final Quality Scores. However, in some instances,

ChatGPT’s translation reads more idiomatic on a sentence level than DeepL’s translation does. When ChatGPT is not translating terms literally word for word from English into German.

4. Interview Results

As mentioned in subchapter 2.2. on methodology, the interviews with study participants were conducted face to face and recorded with a digital voice recorder. Participants also had a piece of paper and a pen in front of them to take notes as needed. Originally, nine interviews were planned in total, but due to time and resource constraints, a total of five were ultimately conducted. The longest interview was 26 minutes long, the shortest 13 minutes and 23 seconds. All interviewees were employed by the Austrian Armed Forces' Language Institute at the time the interviews took place, and all of them worked with English, either in translations or as language teachers. The interview recordings were transcribed using Microsoft Word's text-to-speech function and afterwards edited in those places where the text-to-speech function did not process words or phrases correctly. Apart from a question regarding the intervals of participants' contact with translation as part of their daily tasks at the Language Institute and whether they had been trained as a translator or not, no other personal information was gathered. Interviewees were anonymised and numbered 1 through 5. The target texts that made up the main part of the experiment were printed out and numbered A through C so as to not give the interviewees any indication of the texts' origins.

As previously mentioned in subchapter 2.2., the interview itself was divided into three parts. The first part focused on the interviewee's experience with translation in general and machine translation in particular, as well as their impressions of those machine translations. The second part constituted the experiment itself, within which the interviewees were presented with the three translations without prior knowledge of their origins and given time to look at the texts. Whether they focussed solely on analysing the texts or were talking and already bringing up some of their findings as they went through the texts, depended on the individual. Once they were ready, interviewees were asked to give their ranking of the texts and the reasons for their answers. Depending on how much they found, this part took longer as they went through each of the paragraphs and each of the texts. In part three, participants were then given the actual origins of the texts to see whether their ranking matched the reality. Then, they were asked whether the new knowledge of the texts' origins changed their ranking and their perception of the texts in any way, and about their overall impression of the translations.

The interviews were conducted in German as the target language for the translations was German. The transcripts, as well as the questions, were translated from German into

English by myself for easier understanding and quotation within this thesis. However, the original German transcripts can be found in Appendix A, as well as the full translations and the translations of the interview transcripts. Some of the interviewee's answers were then collected in a spreadsheet to get a statistical overview over all five interviews. Other answers or particular reasons will be quoted were necessary. The interviewee's answers will also be compared to the results of the translation evaluation in chapter 3.

4.1. Translation experience and training

4.1.1. Formal training and translation as part of daily tasks

In order for me to be able to categorise the participants and to see whether their previous translation experience plays a role in their ability to recognise a human or machine translation, it was necessary to ask them about the frequency that they are in contact with translations as part of their daily tasks. This was done with the first question, "Inwiefern haben Sie als Teil Ihrer Tätigkeit innerhalb des ÖBH mit Übersetzungen von Texten zu tun?" ("To what extent are you involved in translating texts as part of your work within the AAF?"). Although there was no explicit question about participants' formal training, all of them volunteered that information on their own.

Out of the five participants, one answered that they had a degree in translation and were trained as a translator but had not officially been hired as a translator at first (Interviewee 3). Interviewee 3 also answered that they had already translated some texts in their time as an intern, before they officially moved onto a position as translator. In between that time, they held another position within the Language Institute but had continuously done translations. The other four participants all answered that they had, or are in the process of completing, a degree in teaching, particularly teaching English.

Two out of five participants answered that they often translate texts or edit translations and act as the second pair of eyes. Although Interviewee 2 answered with the caveat that they primarily coordinate translations and assign them to the right personnel. Three out of the five participants said that they come into contact with translations "now and again". Interviewee 5 said: "Manchmal bitten Kollegen und Kolleginnen mich um Hilfe, weil sie ein zweites oder drittes Augenpaar sozusagen brauchen. Sonst nicht bis dahin." (Interviewee 5: 4) ("Sometimes, colleagues ask me for help because they need a second or third pair of eyes, so to speak. Otherwise, not so much.") It appears that, although, four out of five participants had

not received any formal training as translators, their experience and training as English teachers plays a part in whether or not they are often asked to either assist with editing of translations or take on translations as their own task. This is likely due to a high volume of translation requests and the need for the two-men rule for editing of translations.

4.1.2. Experience with machine translation

The second question in the interview concerned participants' experience with machine translation, specifically ChatGPT and DeepL. The question was "Haben Sie bereits einmal ChatGPT, DeepL oder ein anderes Tool (z. B. Google Translator) zur Übersetzung genutzt?" ("Have you ever used ChatGPT, DeepL or another tool (e.g. Google Translator) for translation?"). I wanted to leave the usage of other machine translation tools open for the participants, even though the thesis focuses on ChatGPT and DeepL, because there are other tools apart from these two and it was likely, the participants had used other tools not mentioned in the question.

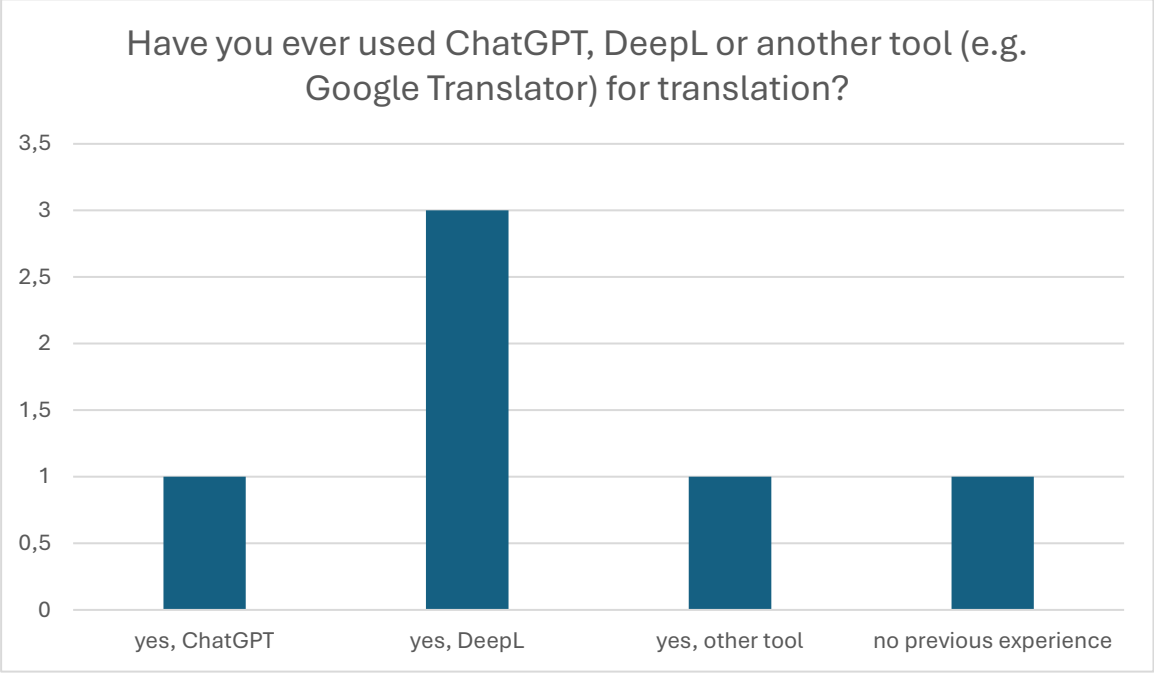

Figure 8: bar chart of participants' answers regarding previous use of machine translation tools

Three out of five participants answered that they had used DeepL before. Interviewee 4 answered that they had briefly used ChatGPT for translation before, but they preferred DeepL. "ChatGPT verwende ich nicht zur Übersetzung so gerne eigentlich. Ich hab es zwar auch schon hier und da gemacht, weil ich im Tool drinnen war, aber wenn Übersetzung, nehm ich lieber DeepL." (Interviewee 4: 10) ("I don't really like using ChatGPT for translation. I've

done it here and there because I was already interacting with the tool, but when it comes to translation, I prefer to use DeepL.") Interviewee 4 also stated that they had recently tried out the new text editing function, DeepL Write, but did not elaborate on their impressions of that tool. Their scant, previous use of ChatGPT as a translation tool, is reflected in the bar chart above, although they said, they preferred to use DeepL. Interviewee 5 said that they frequently used Google Translator for translation but had not previously used DeepL or ChatGPT. And Interviewee 2 said, they had no prior experience with machine translation at all. They expanded on their answer with the following reasoning:

> *Es kann in anderen Settings vorkommen, dass man zur Textinhaltserschließung grobe maschinelle Übersetzung darüber laufen lässt, wenn es nicht geheim ist. Aber das ist nicht unser Anwendungsgebiet. (Interviewee 2: 16)*
>
> *("In other settings, it is possible to run the text through machine translation to deduce the content and meaning of the text if it is not secret. But that's not our area of application.")*

Interviewee 2 was the only one to cite security concerns and levels of security clearance as a reason not to use machine translation on a text. They also said that usage of machine translation for so-called gist translation, the act of roughly translating a text to quickly make its content and meaning accessible, might have been a practice in the past, but they do not use it at the moment. There was also a follow-up question for participants who denied having previously used machine translation. As Interviewee 2 had also given their reasons, i. e. security concerns, differing levels of security clearance of texts, this question was not explicitly asked during the interview.

Apart from Interviewee 4, all other participants denied ever having used ChatGPT before and even expressed surprise that translation was possible with ChatGPT. Most notably, Interviewee 2 answered that they had not really used ChatGPT before and found it interesting that translating was possible with the tool. Although they were aware that DeepL had been explicitly trained and programmed for the purpose of machine translation. Interviewee 5 was the only person to say that they use Google Translator for language combinations other than German into English or English into German. Some of the languages they brought up as source languages were French, Italian and Russian. (Interviewee 5: 11)

4.1.3. Reasons for use of machine translation

The follow-up question to participants' previous use of machine translations was the following: "Wenn ja, warum haben Sie sich für eine maschinelle Übersetzung entschieden?"

("If so, why did you decide to use machine translation?"). This was asked, so as to also get an impression of participants reasons for the use of machine translation. Three out of four answered that the use of machine translation and subsequent post-editing saved them time. Interviewee 4 cited the speed of DeepL and subsequent saving of time as their main reason to use machine translation when appropriate. Interviewee 5 answered that they primarily used Google Translator for gist translation to understand news from foreign countries in languages they did not speak well or did not understand at all. As mentioned in the previous paragraph and subchapter, Interviewee 5 listed French, Italian and Russian as examples for source languages. Italian and French, for Interviewee 5, were examples of "Sprachen, die ich noch so halb beherrsche" (Interviewee 5: 11). ("Languages that I'm only half fluent in.")

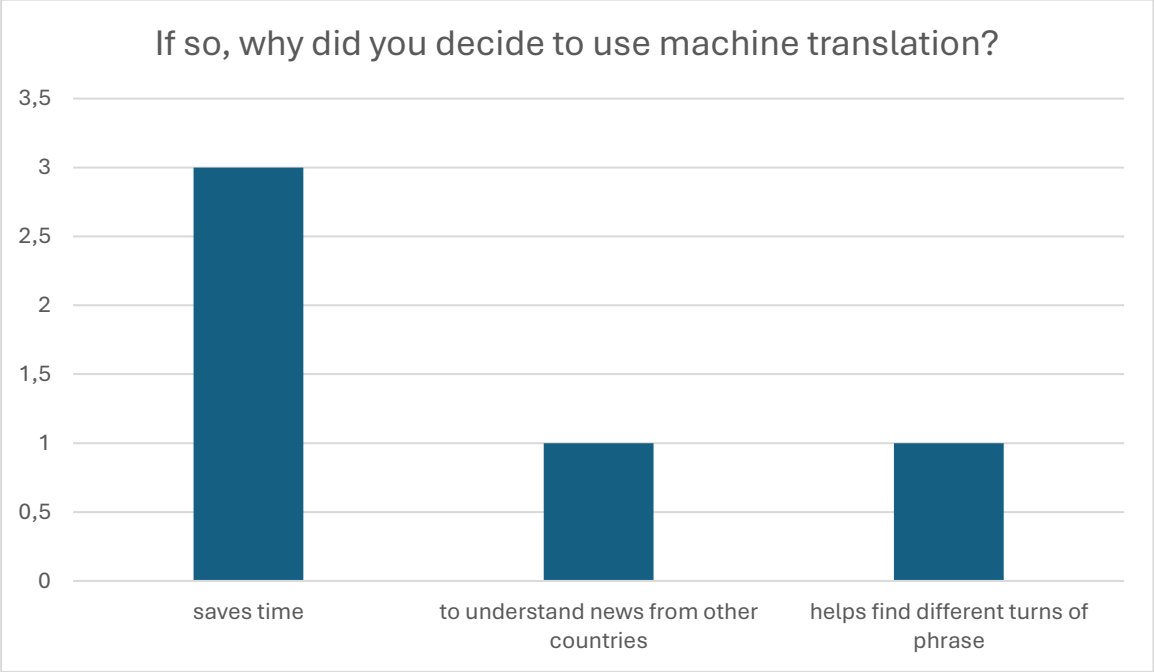

Figure 9: bar chart of responses when asked for reasons for use of machine translation

Interviewee 1 cited the saving of time as a reason, but also said, using DeepL for translation helped them find turns of phrases when they were stuck or also helped them find alternatives when they did not like DeepL's output. As Interviewee 2 did not have experience with the use of machine translation as part of their job, the question was skipped in their interview. But Interviewee 1's answer and them giving two reasons instead of just one, results in the answer total of five, instead of four.

Interviewee 5's answer of wanting to understand news from other countries in languages they did not speak, or did not speak well, aligns with the fictional motive for a translation of NATO's website from English into German or any of the other available three

languages on the website. The text's ultimate function is to communicate. And not every German native speaker speaks English well or possesses the background knowledge necessary to understand some of the domain specifics. NATO's desire to communicate what it is doing for Ukraine and other tasks is already evident in the fact that the website on assistance provided to Ukraine and NATO's reaction to Russia's attack has been translated into Ukrainian and Russian. Due to the fact that French is also one of the alliance's working languages, the existence of a French translation is easy to understand. Why there is no translation into German, if the text's function is communication and dissemination of information, is not. There are a variety of factors that could have played into NATO's decision not to have a German translation. Cost, time and availability of translators are factors, as well as NATO's primary decision on which languages were chosen as target languages, perhaps because it is assumed that German native speakers speak English or French – or Ukrainian or Russian – well enough to understand the gist of the site's content. However, the use of machine translation and subsequent post-editing by a professional German translator with the necessary knowledge and access to resources could be a viable option to mitigate the problem of cost.

4.1.4. Impressions of previous MT use

Another follow-up question to whether participants had used machine translation before was the question on their impression of the translations they had previously seen. The question was phrased as "Wenn ja, was war Ihr Eindruck von der Übersetzung?" ("If so, what was your impression of the translation?"). This was done in order to gather data on how participants had previously evaluated machine translation output they had come into contact with and produced themselves.

All four of the participants who had answered that they had used machine translation before, said that their impression was pretty good, but that post-editing was still necessary to produce a high-quality translation. Interviewee 3 also said,

> *Es kommt darauf an, also wenn es nur so allgemeinsprachliche Sachen sind, dann ist es sehr gut, aber alles was dann irgendwie so ins Militärische mehr geht, das kann DeepL halt von der Terminologie her einfach nicht. (Interviewee 3: 10)*
>
> *("It depends, so if it's just general language things, then it's very good, but anything that goes more into military terms, DeepL simply can't do that in terms of terminology.")*

This view lines up with the results from the evaluation that saw DeepL with three Neutral Terminology Errors and one Minor Terminology Error and ChatGPT with one Neutral, two Minor and 1 Major Terminology Error. But the models' lacklustre performance in terminology can easily be explained by a highly probable lack of terminology resources for the military domain. Texts can be classified, terminology not previously defined and some target languages are not as resource heavy and privileged in research and in bilingual corpora as German and English. Additionally, as has been established during the expert interview, many military terms do not have equivalents in German as the Austrian Armed Forces simply use the English term in their daily life.

4.2. Experiment and remarks from participants

4.2.1. Participants' ranking of target texts

As part of the experiment, participants were asked to do their own ranking of the three target texts in which they assessed which of the texts was the human translation, DeepL or ChatGPT. Participants were therefor given the three target texts, printed out on paper, and numbered A, B and C, as well as a pen and a piece of paper to take notes on. There was no limit on the time given to participants for them to read through the texts, make comparisons and their assessment. Interviewee 5 was the only participant to quickly come to a conclusion, as "ich habe sehr wenige, sehr, sehr wenige Hinweise nur für menschliche Kreativität" (Interviewee 5: 30). ("I have [found] very few, very, very few hints for human creativity")

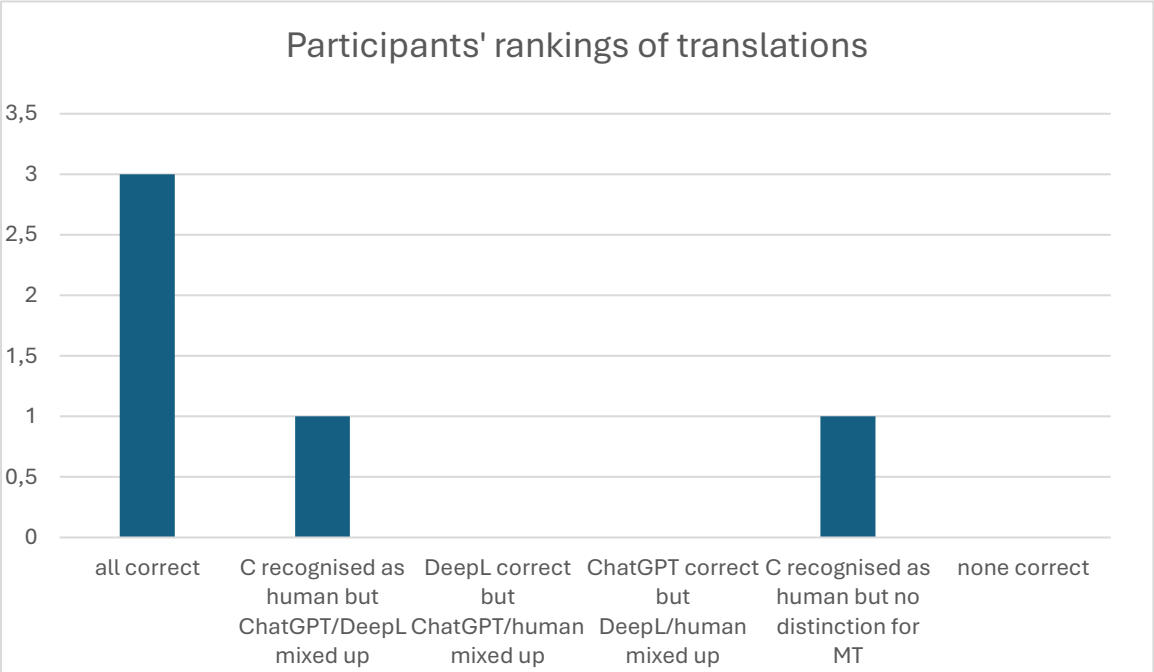

Figure 10: bar chart of participants' rankings as compared with the actual texts' origins

As figure 10 shows, three out of five participants correctly assessed text A as DeepL, text B as ChatGPT and text C as the human translation. Two recognised Text C as the human translation but one had categorised Text A as ChatGPT and text B as the DeepL translation and one had been unable to make any distinction between ChatGPT and DeepL. Out of the three participants who correctly assigned the texts their origins, all three participants possessed previous experience with DeepL, but only Interviewee 4 had previously used ChatGPT for translation as well, if only marginally by their assessment. Of the two participants who had recognised Text C as the human translation but had incorrectly, or not at all, assigned texts A and B to DeepL or ChatGPT, one had answered that they had no or very little experience with machine translation and one, Interviewee 5, had previously answered that they primarily used Google Translator and not one of the two tools used for the experiment.

There appears to be a correlation between participants' experience with machine translation, especially with DeepL, and their ability to distinguish DeepL's translation from other machine translation output. Although, Interviewee 1, for example, confessed that "[…] sicher bin ich mir auch nicht, aber das ist was ich jetzt aus dem schließe" (Interviewee 1: 91) ("[…] I'm not sure either, but that is what I have concluded.") as they knew DeepL's translation quality and were familiar with the tool. The two participants who had been unable to make a distinction between DeepL and ChatGPT's outputs, had little to no prior experience with DeepL and ChatGPT. However, this correlation could also come down to participants' general experience with translation and how often they translated or edited a translation.

4.2.2. Reasons for Ranking

After the participants had presented their rankings, they were then asked to give reasons and examples from the texts as to why they had chosen to assign them a certain label. In this chapter, I will go through the reasons given by each participant one by one and afterwards attempt to align those reasons with the errors and other hints found by myself during the evaluation of the translations.

Interviewee 1's first reason was, text C "hat für mich passendere Formulierungen" (Interviewee 1: 57). ("has more appropriate wording in my eyes") They then went on to list examples. The first example they gave was "NATO and Allies" in paragraph 2 of the source text and all translations. ChatGPT translated this in text B as "NATO und Alliierte", and translator T wrote "die NATO und ihre Verbündeten" in text C as did DeepL for text A. To

Interviewee 1, the translation choice made for text C was more logical, “weil da geht es ja nicht um die Alliierten, sondern um Verbündete von der NATO” (Interviewee 1: 61) (“because it's not about the Allies, but about NATO allies”) as “Alliierten” in the DACH region is strongly connected to the Allied powers of World War 2, so the United Kingdom, the Soviet Union, the United States and China. I found the same error during the evaluation part and came to the same conclusion. As the target audience for the fictional translation commission of the experiment is Austrians, particularly Austrians enlisted in or employed by the Austrian Armed Forces, using “Alliierte” for allies when the context does not permit for the reader or the translator or author of the text to make a connection to World War 2 would cause unnecessary confusion for the target audience.

Another example Interviewee 1 gave was “Command, Control, Communications and Computers (C4)” in paragraph 4 of all texts. Interviewee 1 said, “Es wurde im Text C als Führung übersetzt, was ich, glaube ich, passender finde als Kommando” (Interviewee 1: 69). (“In text C it was translated as ‘Führung’ which I feel is more appropriate than ‘Kommando’.”) They also said, they thought that text C’s translation choice, “Führung”, was probably translated with background knowledge in mind which the machine translation models did not possess. This tracks with translator T’s professional background and their access to resources such as dictionaries and glossaries published by the Austrian Armed Forces’ Language Institute.

The next example Interviewee 1 gave was “servicemen and women” in paragraph 5, which was translated in texts A and C as “Soldatinnen und Soldaten”. ChatGPT translated it as “Dienstmänner und -frauen”. Interviewee 1 explained text B’s choice as “[w]as eigentlich nicht gesagt wird” (Interviewee 1: 75). (“[w]hich is not usually used as a term”) The actual use of ChatGPT’s translation choice was explained in Chapter 3.3 as a vassal or a porter. The correct translation in the given context was “Soldatinnen und Soldaten” as was pointed out by Interviewee 1 and done by DeepL and translator T. However, it is highly likely that DeepL possesses more training data where “servicemen and women” has been translated and aligned with “Soldatinnen und Soldaten”. ChatGPT likely does not have that same training data.

Interviewee 1 also cited DeepL and ChatGPT’s translation choices of “improvisierten Sprengsätzen” in the last paragraph for the original “improvised explosive devices” as “Das klingt für mich ein bisschen fremd” (Interviewee 1: 81). (“That sounds a bit strange to me.”) They went on to say, that they felt the meaning of “improvised” was slightly different to the

German “improvisiert” and they gravitated towards the choice of “behelfsmäßiger Spreng- und Brandvorrichtungen”.

For the term “small arms and light weapons”, Interviewee 1 again gravitated towards Text C’s phrasing of “Klein- und Leichtwaffen” in the last paragraph as it “klingt einfacher und […] stimmiger” (Interviewee 1: 88). (“reads more easily and […] feels more appropriate”) Although both the version in Text C as well as the translations in texts A and B, “kleine und leichte Waffen” are used in German as previously mentioned in chapter 3.2. on the evaluation of the DeepL translation.

Interviewee 1 also noticed that the machine translations had two spaces instead of one in some places where the human translation did not. They also specifically looked for cases of gender-equitable language use. They said, it was to see “Ob da vielleicht irgendwelche Clues, irgendwelche Hinweise darauf sind, was von wem übersetzt worden sind. Aber ich glaube, teilweise wurde gegendert und teilweise nicht” (Interviewee 1: 109). (“if there are any clues, any hints as to which text was translated by whom. But, I think, sometimes gender-equitable language was used and sometimes not”) The inconsistent use of gender-equitable language then serves as a marker that confused the participant as it was not consistent in the human translation either.

Interviewee 2 gave their reason for choosing text C as the human translation, “Weil die ist flüssig, lesbar und weil ich davon ausgehe, dass die Person auch recherchieren kann und nicht einfach irgendwie… Also auch tatsächlich die richtigen Quellen verwendet, um zu recherchieren” (“Because it is fluent, readable and because I assume that the person can also do research and not just somehow... So actually uses the right sources to do research”). Thus, while Interviewee 1 primarily cited audience appropriateness and terminology as their reasons for their ranking, Interviewee 2 focused more on fluency and readability. Interviewee 2 also said, they had previously not worked with ChatGPT a lot and did not know that it could be used for machine translation. They also said, because they did not know that ChatGPT could be used for machine translation, it was difficult to distinguish between DeepL and ChatGPT for the machine translated texts. Their assumption was that ChatGPT would be better because “ChatGPT über mehr AI-Kapazität verfügt, Dinge herauszufinden ist es vielleicht… Kann vielleicht ChatGPT besser die richtigen Begriffe zuordnen?” (Interviewee 2: 49) (“ChatGPT has more AI capacity to figure things out, maybe it's... Maybe ChatGPT is better at matching the right terms?”)

Interviewee 2 also noted some of the terminology as clues to the origin of the translated texts. “Behelfsmäßige Spreng- und Brandvorrichtungen” as it is found in the last paragraph of text C was a big clue for Interviewee 2 because, to them, it is a typically Austrian choice of wording. “Also ein typisch österreichisches Wording, das man wahrscheinlich nur verwenden kann, wenn man, wenn man die österreichischen Vorschriften kennt, beziehungsweise Fachliteratur, also österreichischer Provenienz verfügt.” (Interviewee 2: 54) (“This is typical Austrian wording that you can probably only use if you are familiar with Austrian regulations or specialised literature, i.e. of Austrian provenance.”) Interviewee 2 then moved on and cited ChatGPT’s translation “Dienstmänner und -frauen” in paragraph 5 as an example for “[u]nglückliche Benennungen”. (Interviewee 2: 56) (“unfortunate choice of words”).

> *Ich weiß es nicht ob DeepL oder ChatGPT besser in der Lage ist abzugreifen, dass natürlich „service personnel“ nicht „Dienstmänner“ sind. Also, nachdem ich das nicht genau weiß, war das ein bisschen ein Unsicherheitsfaktor für mich, aber ich vermute trotzdem, dass B DeepL ist und A ChatGPT. (Interviewee 2: 58)*
>
> *(“I don't know if DeepL or ChatGPT are better able to pick up that of course ‘service personnel’ are not ‘Dienstmänner’. So, not knowing that for sure, that was a bit of an uncertainty factor for me, but I'm still guessing that B is DeepL and A is ChatGPT.”)*

Interviewee 2 did not notice any inconsistencies with gender-equitable language in any of the texts. And them mentioning that their lack of experience with ChatGPT translations resulted in a feeling of uncertainty when it came to deciding which of the machine translated texts was done by ChatGPT and DeepL. They also made the interesting assumption that ChatGPT would be better at matching the right terms in German to the English originals because of its “AI capacity” despite the fact that ChatGPT, as a Large Language Model was not specifically designed to process translation tasks. They also posed the question whether ChatGPT was “noch weiterentwickelt” (Interviewee 2: 59) (“further developed”) and thus able to do its own research into the correct terminology for a domain-specific text as was used in the experiment.

Interviewee 3’s first example for why they chose text C as the human translation was a point Interviewee 1 also mentioned, the translation of “Command, Control, Communication and Computers” in paragraph 4 of the texts. To Interviewee 3, the translation of the term as “Führung, Information, Kommunikation, Computersysteme” was the first clue that solidified their choice for text C as the human translation, “weil es ist halt Command and Control und das ist einfach wörtlich übersetzt und das übersetzt man aber eigentlich nicht so wörtlich, wenn man Hintergrundwissen hat” (Interviewee 3: 21) (“because it’s Command and Control

and that is just a literal word-for-word translation and nobody translates it as literal if they have the background knowledge."). Interviewee 3 also focused on giving examples of terminology to illustrate their choice and to defend their ranking, but they also mentioned fluency, sentence structure and wording.

Making a distinction between ChatGPT and DeepL was still a difficult task for Interviewee 3 as they said:

> *Also ich hab jetzt voll lange überlegt, weil da einmal ist es als Soldatinnen und Soldaten und einmal ist es als Dienstmänner und -frauen und wahrscheinlich ist es auf Englisch einfach service men. Und das ist halt auch schon wieder so komplett wörtlich übersetzt. Und da war ich mir halt nicht sicher, ob ChatGPT so vielleicht so wörtlich übersetzt, weil ich eher glaub, dass DeepL das schon als Soldaten, aber ich bin mir nicht sicher, keine Ahnung. (Interviewee 3: 21)*
>
> *("So I've been thinking about it for a long time now, because one time it's translated as 'Soldatinnen und Soldaten' and one time it's translated as 'Dienstmänner und -frauen' and it's probably just 'service men' in English. And that's also a completely literal translation. And I wasn't sure whether ChatGPT might translate it so literally, because I think that DeepL translates it as 'Soldaten', but I'm not sure, I don't know.")*

Again, their lack of experience with translations with ChatGPT caused some unease in their decision-making as Interviewee 3 was unsure how ChatGPT might handle some terms during translation like "service men", but then they also focused on sentence structure and grammar to distinguish the two machine translations. Although Interviewee 3 also said, they noticed the same or similar sentence structures in the human translation, which lead them to also believe that the human translation might have been done with the help of DeepL. As I had given translator T no explicit instructions not to use machine translation, it seems likely that they did. Interviewee 3 also noticed that the machine translations had problems with keeping to the source text style with the bulleted lists which, according to them, was another clue that texts A and B were machine translations.

Interviewee 3 then went on to explain that, despite DeepL and ChatGPT not always using the correct terminology – apart from the "servicemen and women" which turned into "Dienstmänner und -frauen" for ChatGPT -, their terminology errors did not usually change the meaning too much. To Interviewee 3, the texts were still comprehensible, and the meaning was still there despite those terminology errors, which, they said, are all terms that only an experienced translator or a translator with access to the right terminology resources could have translated correctly. Difficult terminology was one of the reasons why I chose the text,

and this piece of text in particular, as I had assumed that it would pose a challenge for both ChatGPT and DeepL.

Interviewee 4 made their assessment after reading only the first paragraph of the three texts. They said that text A and text B were very similar with the differences, again, coming down to terminology, and they also mentioned similarities in sentence structure between text A and text C. They did later elaborate and analyse all paragraphs of the texts. Despite their confidence in declaring text C the human translation, which was correct, they were unsure about making the distinction between DeepL and ChatGPT. Like Interviewee 2, they came to the assumption that "[i]ch glaub ChatGPT hat inzwischen die Fähigkeit, kann man das so sagen, auch Synonyme eher sinnvoller einzusetzen. Und auch nicht so am Wort zu kleben wie DeepL das auch tut" (Interviewee 4: 20). ("I think ChatGPT now has the ability, can you say that, to use synonyms more sensibly. And not to stick to the word like DeepL does.")

They went further to talk about the instance of "best practices" in the first paragraph, for which they said, there is no real equivalent in German. "Und da schreiben sie Praktiken, und das hat offensichtlich der menschliche Übersetzer mit Best Practices übersetzt, was man ja auch eigentlich eher verwenden würde, weil Praktiken ist in diesem Zusammenhang sinnlos." (Interviewee 4: 18) ("And there they write 'Praktiken', and the human translator has obviously translated that as best practices, which is what you would actually rather use, because 'Praktiken' is meaningless in this context.") This was also noted during the translation evaluation and the expert interview conducted with translator T.

Another example brought up by Interviewee 4 for their thinking process in the ranking was "reorganising" in paragraph 5, which DeepL and ChatGPT both translated as "Reorganisation" and translator T translated as "Umstrukturierung". "Umstrukturierung ist eher ein Wort, das wir verwenden würden vielleicht im Deutschen als Reorganisation." (Interviewee 4: 33) ("'Umstrukturierung' is more a word that we would perhaps use in German than "Reorganisation".)

They, like Interviewee 1, also noted the lack of gender-equitable language and the inconsistent use of it in German. They said, "Ich glaube auch, dass ein Mensch 'Mitarbeiter und Mitarbeiterinnen' schreiben würde. Hoffentlich." (Interviewee 4: 34) ("I also believe that a human would write 'Mitarbeiter und Mitarbeiterinnen'. I hope."). And for the civilians in paragraph 7, which all three texts translated as "Zivilisten", they suggested "Zivilbevölkerung" to keep it gender-neutral.

For Interviewee 5, “best practices” in the first paragraph was the first clue to text C’s origin. As they pointed out, “Es ist ja nicht abnormal englische Begriffe, englische Idioms in deutschen Texten anzutreffen. […] Wohingegen der Rechner oder die KI alles zu übersetzen pflegt.” (Interviewee 5: 26) (“It is not abnormal to find English terms, English idioms in German texts. [...] Whereas the computer or AI tends to translate everything.”) Interviewee 3 also pointed out the machine translation models’ tendency to translate terms literally rather than with the correct domain-specific terminology and to apply the source language grammar and syntax, in this case English, to the target language, German. While this occurrence can easily be explained with a lack of the right texts within the training data and the fact that neither DeepL nor ChatGPT were specifically designed to translate texts in a domain like the military, both Interviewees brought those occurrences up separately. Interviewee 4 did not specifically say that DeepL and ChatGPT translate terms literally, but they also said that DeepL had a tendency to stick to the English syntax and apply that to the German target text.

Another thing Interviewee 5 pointed out was the three different translations of “assistance package” with text A listing “Beistandspaket”, text B “Hilfspaket” and text C “Unterstützungspaket”. They did not, however, elaborate on this other than that they found it interesting. They also noted small differences in meaning for some of the terminology like “Bündnis” and “Allianz” for “alliance” in the first paragraph, but assumed that the machine translation models, and ChatGPT in particular, would use “Bündnis” more than “Allianz”. About the quality of the translations, they said, “ich sehe qualitätsmäßig, qualitätsmäßig wirklich, wirklich sehr, sehr wenig Unterschiede. Wenn ich ehrlich bin.” (Interviewee 5: 37) (“I really, really, really see very, very little difference in terms of quality. If I'm honest.”).

Overall, as figure 11 shows below, the reasons for participants’ rankings can be classified as follows. All five of them cited terminology as either clues or as diversions for their ranking. Two, Interviewee 2 and 3, cited fluency and readability of the target texts, with the caveat that both DeepL and ChatGPT also created very fluent texts where the fluency distracted from errors such as underlying issues with terminology or even the sentence structure. Interviewee 3 and Interviewee 4 had both mentioned machine translation models’ tendency to stick to the source language’s sentence structure and simply copy-paste the target language’s words onto that structure. Audience appropriateness and wording, or style, were mentioned by two and four participants respectively. Audience appropriateness considered things like the translation of “allies” in paragraph 1 as “Verbündete” instead of “Alliierte”. And two people, Interviewee 1 and Interviewee 3, noted formatting as part of their reasoning.

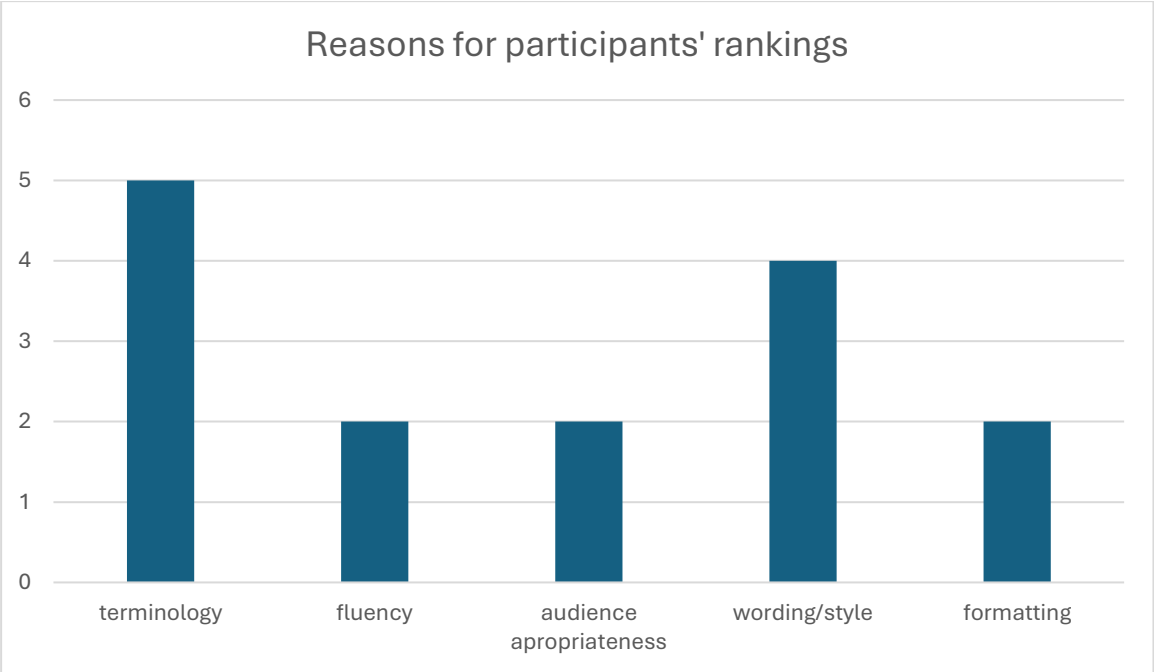

Figure 11: bar chart of categories of reasons for participants' rankings

Participants were not explicitly asked about their confidence in their rankings, but they themselves provided clues as to how they felt. All of the participants said once or twice that they were pretty confident about their ranking of text C as the human translation but felt unsure as to how to differentiate between the machine translations.

4.2.3. Further remarks on participants' rankings

After being told the actual origins of the three target texts, participants were then asked if they had any further remarks about their rankings and the machine translations in general. Not all participants had final remarks to make, but all agreed that the quality of the raw machine translation outputs surprised them and all agreed that both machine translations could be brought to the same level in quality as the human translation with post-editing done by a translator or post-editor with the necessary background knowledge.

As mentioned in the previous subchapter, Interviewee 1 also noted the formatting and some extra spaces between words with the machine translated texts that should not be there. They also said, they weren't sure, why those spaces were there or if it was an error within the machine translation model. After the reveal of the texts' origins, they said, they were surprised but also happy they had "guessed" right. Afterwards, when they talked about the use of gender-equitable language again, they said the following: "Ja, interessant. Also ich glaube im Text B steht dann auch Patienten, warum dann Soldatinnen und Soldaten übersetzt wird, aber nicht Patientinnen und Patienten, aber es hat also das hat … von dem aus habe ich nicht

schließen können." (Interviewee 1: 113) ("Yes, interesting. So I think in text B it also says 'Patienten', so why is it translated 'Soldatinnen und Soldaten' but not 'Patientinnen und Patienten', but it has, so that has ... I couldn't deduce from that.") I then also mentioned that I had specifically looked at the Ministry of Defence's regulation regarding the use of gender-equitable language and explained the steps in there, that authors could choose whether to use simply the generic masculine with a disclaimer at the beginning of the text or to write out the masculine and feminine versions of words. Interviewee 1 then went on to say that "[i]st interessant, dass sie sich in dem Punkt, die drei, die Maschine und der Mensch, einig sind." (Interviewee 1: 118) ("It's interesting that the three of them, the machine and the human, agree on that point.") They meant "agree" here in the sense that all three texts did not employ a consistent strategy for the use of gender-equitable language.

Interviewee 2 first made final remarks on their ranking and their mistake of assigning text A the label ChatGPT and text B the label DeepL. "Und dann war auch sozusagen meine Vermutung falsch. Dass ChatGPT, auch Kontextbezogen, die Terminologie rausfinden kann, ist DeepL besser, in diesem Fall klar." (Interviewee 2: 64) ("And then my guess was also wrong, so to speak. The fact that ChatGPT can find out the terminology, also contextually, DeepL is better, in this case clearly.") As previously mentioned, they had ranked text A as ChatGPT and text B as DeepL under the assumption that ChatGPT's Transformer and its "AI" was better able to research terminology and leverage its inherent technology to solve translation problems. They also had the following to say about the quality of all three texts: "Wenn man, wenn man das mit Post-editing macht, dann ist das, dann ist das ja wirklich faszinierend, wie hoch die Qualität ist. Aber die Humanübersetzung ist trotzdem besser." (Interviewee 2: 77) ("If you do that with post-editing, then it's really fascinating how high the quality is. But the human translation is still better.") But they, again, highlighted that it is impossible for the Language Institute to just use DeepL or ChatGPT on certain texts because of security concerns and different classification levels for the texts.

Interviewee 2 then also spoke about the intentions behind the NATO website and the strategies and policies involved in the translation of that website as there are official translation for French, Russian and Ukrainian. They said, "Mit denen möchte man natürlich auch kommunizieren. Die NATO ist natürlich interessiert daran, dass die Öffentlichkeit in Russland und in der Ukraine über das Bescheid weiß, ist klar." (Interviewee 2: 95) ("Of course, you also want to communicate with them. NATO is naturally interested in the public in Russia and Ukraine knowing about this, of course.") This also ties back into the source text

analysis in Chapter 2 where I also laid out the motive and the function of the source text as informative and to aid communication with the public on NATO's activities and tasks.

Interviewee 3 elaborated on their argument that machine translation models tend to stick too closely to the source text's and source language's sentence structure by saying that

> *Ja, also ich find die sind immer so umständlich. Die Sätze, die bleiben dann immer in so einer Satzstruktur, die wenn… Wenn das ein Mensch sagen würd, man würd es nicht so komplett verschachteln, sondern man würd versuchen, dass man das irgendwie in eine schönere Form bringt. Und die maschinellen Übersetzungen, finde ich, bleiben immer so. (Interviewee 3: 27)*
>
> *("Yes, I think they're always so awkward. The sentences always remain in a sentence structure that if... If a human were to say it, it wouldn't be so completely convoluted, but you'd try to put it into a nicer form somehow. And I think machine translations always stay like that.")*

To Interviewee 3, machine translation models are incapable of adapting a translation on their own to aid readability and also correct the sentence structure on their own after translation. However, there are two caveats: First of all, would this then still be counted as raw machine output or would it have already gone through a stage of post-editing – even if that was done by the model itself. And second, whether ChatGPT would be able to parse its own translation output and edit it like a human if the explicit instructions were given in the initial prompt would need to be researched. Interviewee 3 is right in arguing that a raw machine translation without post-editing is acceptable for gist translations where the content is more important, but that post-editing would definitely need to be done by a skilled post-editor or translator with the right knowledge background.

Interviewee 3 also talked about the differences in thought process and the way translators read as compared to the, perhaps monolingual, target audience who might not possess the same training and look for mainly only the information in a text.

> *Und ich finde, gerade wenn du als Mensch übersetzt, denkst du ja eher noch, du denkst da noch anders darüber nach, als wenn du das einfach nur liest. Wenn du das jetzt einfach so auf Deutsch liest, erfasst du halt die Information und denkst dir ja, passt. Aber wenn du es übersetzen musst, dann liest du es und wenn es ein langer Satz ist, dann denkst du nochmal so: okay, was sagt das jetzt wirklich aus? (Interviewee 3: 30)*
>
> *("And I think that especially when you translate as a human being, you tend to think about it differently than when you just read it. If you just read it in German, you simply grasp the information and think to yourself, that's fine. But if you have to translate it, then you read it and if it's a long sentence, then you think again: okay, what does that really say?")*

This difference in reading and processing a text is also why the original plan for the thesis experiment had been to present the translated texts to enlisted and civilian personnel without prior translation or language teaching training to see how they would read the texts and whether they would be able to tell a difference. However, even with the current, small participant pool, there was a lot of variety regarding experience and training.

Interviewee 4's only remark at the end was about the human translation being heavily influenced by DeepL.

> *Außer dass Mensch sehr stark von DeepL beeinflusst ist offensichtlich und wirklich ganz offensichtlich mit dem Programm gearbeitet hat und dass da auch noch Luft nach oben ist, weil gewisse Dinge man trotzdem nicht sagen würde oder schreiben würde. Das ist alles, aber ich habe das Problem in meinem täglichen Umfeld. (Interviewee 4: 45)*
>
> *("Except that the human translation is obviously very strongly influenced by DeepL and the translator has obviously worked with the programme and that there is still room for improvement because you would still not say or write certain things. That's all, but I have the problem in my daily environment.")*

They did allude to the two-man rule as applicable to both machine and human translation – as is also defined by ISO 17100 – and, as has been previously said by both participants and me, the machine translation would definitely benefit from post-editing if it were intended for publication like the fictional translation commission would have detailed, with the source text being a website intended for communication with the public.

Interviewee 5 also insisted on the necessity for a human pair of eyes in their final remarks when they said: "Ich glaub, dass AI sehr wohl und sehr gut für solche Übersetzungen einzusetzen ist. Nur dann, wenn am Ende ein Mensch drüber schaut. Weil der wird tatsächlich sich mit einigen Begrifflichkeiten, und wie die eingesetzt werden, besser auskennen." (Interviewee 5: 43) ("I believe that AI can be used very well for such translations. Only if a human looks over it in the end. Because they will actually be more familiar with some of the terminology and how it is used.")

All participants ultimately agreed that the quality of the raw machine translation output was pretty good, but the texts would benefit highly from post-editing done by skilled editors or translators who are also familiar with the domain. With the Austrian Armed Forces and certain peculiarities of the Austrian German locale, this comes as no surprise. A translator employed by the Bundeswehr or even the United Nations would not translate the text the

same way the human translator did for this experiment. Even if the differences mostly come down to terminology.

5. Analysis of Results and Conclusion

5.1. DeepL and ChatGPT's ability to translate military texts

Research question 1 asked "How well can DeepL and ChatGPT handle translating texts in the military domain from English into German?". Over the course of this thesis, I attempted to answer this question and also investigated how Hypothesis 1 applied to the investigated text. Hypothesis 1 stated that "DeepL can produce a fairly fluent target text in German but will probably mix terminology from the German Bundeswehr with the Austrian Armed Forces and the German locale with the Austrian one, while ChatGPT might be able to disguise terminology issues and distract with issues in style and idiomatic use of language." In this chapter, I will take a closer look at the results from both the translation evaluation and the interviews to see whether Hypothesis 1 was true or false.

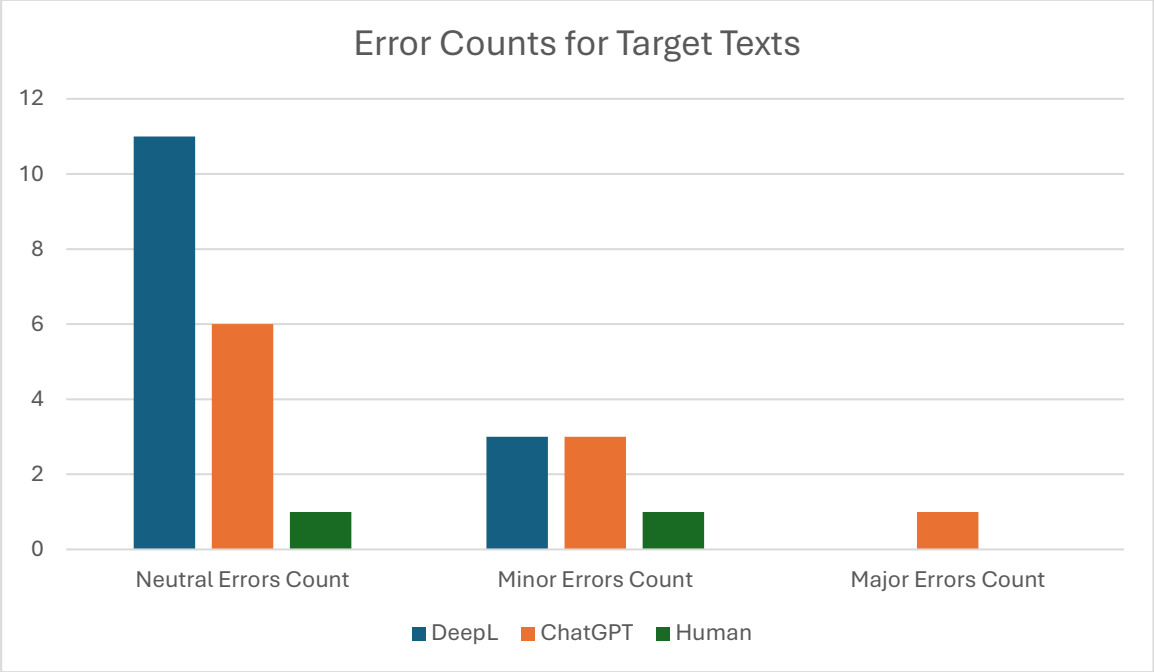

Figure 12: bar chart of errors count found over all three texts during translation evaluation

As figure 12 above shows, DeepL is leading the chart with 11 Neutral Errors in total and is neck and neck with ChatGPT in the Minor Errors Count category. However, as we recall from Chapter 2, DeepL's final Raw Quality Score was 98.662% and the Calibrated Score was at 86.622%. DeepL's Raw Quality Score was 3.6789% higher than ChatGPT's (RQS:

94.9832%), and the Calibrated Quality Score was higher than ChatGPT's by 36.7892% (CQS: 49.8327%). The scores are not the most important metric of measuring the quality of the translated texts in this experiment and how well they did on their given task. While, yes, the Error Type Weight for Terminology Errors was higher than other categories, usage of the right terminology is important for texts within the military domain, as it is for everyday speech in the same domain. While one can explain orders or specific terminology in a face-to-face conversation or if there is space for it in a text, the time is often not there. And, as the interviews have also shown, there is specialised terminology already laid out for the translation of texts within the Austrian Armed Forces. Neither DeepL nor ChatGPT had access to those terminology resources.

While evaluating the quality of the translations, it is also important to consider the context and the text as a whole. While the machine translations did technically fail their evaluation because of too many errors in terminology, as the interviews also showed, the texts are still fluent, and the errors did not change the meaning too drastically. DeepL also came with the caveat that it was not possible to pick the locale for German and DeepL Translator, unlike ChatGPT, also has no option to add specific instructions on how to handle terminology. This is also why I did not give ChatGPT specific instructions for the translation of the terminology as I wanted both to operate from a similar set of instructions.

From purely looking at the Raw and Calibrated Quality Scores and the Quality Rating, the machine translations did not handle the texts well. But, with the MQM even advocating for explicitly training evaluators in how to evaluate and how to use the categories on the scorecard, it is important to look at the whole texts and not just at the score and the final rating that the scorecard gives for the translations. With the first paragraph, for example, there is a translation error as "best practices" would not be translated as "bewährte Praktiken" by a human translator with the knowledge that "bewährte Praktiken" is not commonly used in German and keeping "best practices" would simply be better. The first paragraph, however, is also one sentence that stretches over six lines for DeepL, five lines for ChatGPT and six lines, again, for the human translation. During editing, even if it is not post-editing of a machine translation, the first paragraph would benefit from being split into two sentences. This would not only increase the readability of the text, but also help the translation to move further from the source text. As the source text function is communicative and informative, and it is equally important to keep those functions intact for the target text, a functional translation strategy would benefit the text here. Neither DeepL nor ChatGPT can make these decisions on

their own. Whether DeepL Write would be able to produce a translation with that translation strategy in mind, would require further research.

Holland et al. had also observed the exercises in the late 90s and early 2000s where embedded machine translation had been field tested by the U.S. army. The main problem those studies faced was the OCR component of the system and varying sizes of training data available for the language combinations. According to Holland et al., "Users appreciated the utility of imperfect MT: 'Enough was usually translated to allow for an effective first screening...A real linguist would have to do the actual translation.'" (2000: 245) DeepL and ChatGPT definitely did a lot more than simply translate the text enough to allow for a screening before the text was then given to the human translator. For DeepL and ChatGPT's texts, the terminology was the main problem. But DeepL handled this better, although it has a higher Neutral Errors Count – of which only three were in terminology. Neutral Errors do not count towards Penalty Totals and so do not have an impact on the Quality Scores. It is ultimately up to the evaluator and the person who tasked them with the evaluation to decide whether the translation was really a fail or a pass. For DeepL, the raw machine output is still of such a quality that it is more than acceptable, and with post-editing can even be made better. ChatGPT's strengths do not lie in terminology either, but it does read more idiomatic, for example in paragraph three.

> „Ergänzend zum CAP wurden seit 2014 mehrere Treuhandfonds eingerichtet. Diese Treuhandfonds stellen Mittel zur Unterstützung der Fähigkeitsentwicklung und des nachhaltigen Kapazitätsaufbaus in wichtigen Bereichen bereit." (ChatGPT-4 2024: par. 3)
>
> (Original: "Complementing the CAP, several Trust Funds have been launched since 2014. These Trust Funds provide resources to support capability development and sustainable capacity-building in key areas.")

Although "Kapazitätsaufbau" is still to close to the source text and does not sound very idiomatic, the rest of those two sentences are definitely acceptable for a first draft translation. As Interviewee 4 had noted, the human translation also possesses certain parts and paragraphs that would have to be edited before publication to make them more accessible to the reader.

All five of the study participants noted how surprised they were at the high quality of the raw machine translation output. Which is why, it should be again noted, that for both DeepL and ChatGPT, if one were to disregard the importance of the terminology and the inconsistent use of gender-equitable language, both texts would pass a quality evaluation. They would not pass without raising some flags about the grammar and the audience appropriateness, as well as the style. But their overall score would look different.

However, with terms like “Allianz” and “Bündnis”, which are both valid translations for “alliance”, context and the target locale matter. NATO, in German, is rarely called an “Allianz”, although by definition, it would be. According to the *Digitales Wörterbuch der deutschen Sprache*, a “Allianz” is a “Bündnis zwischen Staaten” (“an alliance between two or more countries”) (2016). The same applies to the entry for “Bündnis”, but the rate of use is different for those two words. The trajectory in the graph provided by the DWDS clearly shows that “Bündnis”, the pink line in figure 13 below, has always been higher than the line for “Allianz”. (2024a, 2024b) This means that “Bündnis” was in use more in, in this case, newspaper articles monitored since 1946 than “Allianz”, and it then stands to reason that a text from 2022, with a translation done in 2024, would use “Bündnis” when talking about NATO instead of “Allianz”, although it is an alliance in English.

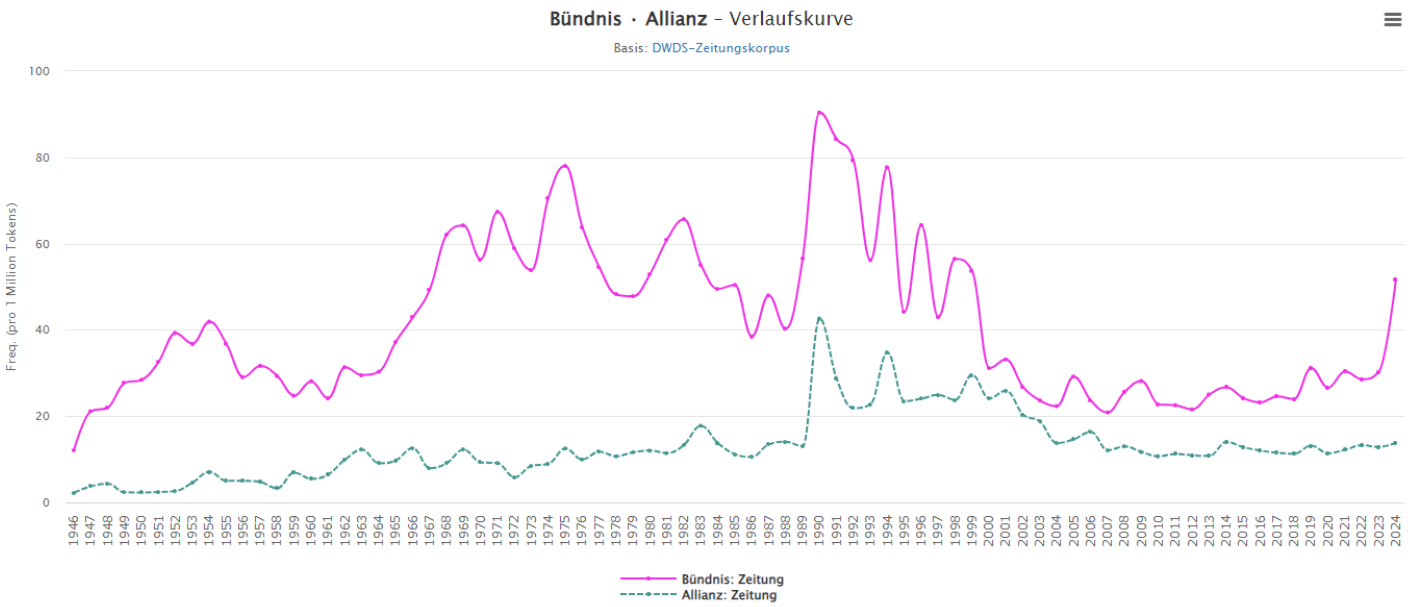

Figure 13: trajectory of use of the words "Bündnis" (pink) and "Allianz" (green) in newspaper articles in German since 1946

“It has been suggested that the field of communication studies needs to undergo a Copernican revolution by acknowledging that AI belies the deep-seated understanding of technology as a non-agentic instrument or means, which had dominated the field for decades (Gunkel 2020, 66–67).“ (Asscher 2023: 2–3) Asscher and Gunkel here argued that AI – and machine translation as part of the field of Artificial Intelligence – have no agency of their own as is often assumed through popular culture. That AI has no place in communication studies because it does not follow a specific goal. However, I would argue, and Asscher does later as well as has been previously mentioned, that machine translation and AI-base machine translation, or Large Language Model translation like it was done with ChatGPT here, while

not following a goal of its own, does attempt and ultimately achieves the goal it was set by the human operator.

The goal for the fictional translation commission was a target text in German that keeps the same text functions as the source text. Both DeepL and ChatGPT's texts do work as pieces of communication. While ChatGPT's use of "Dienstmänner- und frauen" instead of "Soldatinnen und Soldaten" can definitely cause confusion, and warranted the Major Terminology Error type, the texts function well as a whole. Given that they are only the raw machine translation output without post-editing, publishing the texts as is would definitely raise some questions. ChatGPT's text more so than DeepL's. However, it bears repeating that the study participants all had either experience with translation or had been trained and worked as English language trainers for predominantly German native speakers. They would, of course, catch errors that the average person stumbling across NATO's website would likely not.

As Interviewee 5 had mentioned at the end of their interview, "Vielleicht, dass sich das in Zukunft verändert, vielleicht, dass GPT-5 eine noch bessere Arbeit geleistet hätte in diesem Bereich. Aber AI… fantastisch in Kombination mit einem Paar menschlichen Augen. Das würde ich sagen." (Interviewee 5: 43) ("Maybe that will change in the future, maybe that GPT-5 would have done an even better job in this area. But AI... fantastic in combination with a pair of human eyes. That's what I would say.") ChatGPT-4 and the free web version of DeepL Translator were used for this thesis experiment. The results might change in the future as GPT-5 and DeepL Translator develop. Perhaps, there might even be a machine translation engine in the future that has been specifically trained on military texts. Or perhaps NATO might change their policy and require translations into more than their working languages and Ukrainian and Russian.

Ultimately, DeepL handled the source text better in terms of terminology than ChatGPT did. A combination of both texts with a round of full post-editing prior to publication could raise the texts to the same level of quality as the human translation. Whether this would be true for other language combinations, especially in low resource languages, remains a point of further research.

5.2. Study participants' ability to differentiate machine translations from human translation and their perception of machine translation

Apart from an interest in knowing how well machine translation, in particular DeepL and ChatGPT, can handle texts within the military domain, knowing if and how readers within the target audience of those translations can differentiate between not only machine translation and human translation, but also different machine translation models. Or, in ChatGPT's case, LLM translation. This was not only due to wanting to know how other people perceive machine translation, but also because this knowledge could be used in the future to train translators as post-editors and to strengthen a society's resilience against disinformation. While DeepL and ChatGPT did not exhibit hallucinations with the present text and thus one could have argued that they could be used for disinformation campaigns, ChatGPT's other function, like text or image generation, could. As the translation industry seems to shift more towards machine translation and post-editing, it seems more vital from a translator's perspective to develop the means and the skills for translation evaluation and post-editing.

Research question 2 had been formulated as "How do employees of the Austrian Armed Forces perceive and differentiate between translations done by ChatGPT, DeepL and human translations?" As already explained, the original plan for the experiment was to interview enlisted personnel as well as civilian personnel with the Austrian Armed Forces. Due to my inability to establish contact with soldiers and officers, the experiment parameters were changed to just civilian personnel, particularly within the Language Institute at the National Defence Academy in Vienna.

Hypothesis 2, in conjunction with Research Question 2, stated the following:

> "Austrian Armed Forces' employees with no prior experience with ChatGPT, DeepL or translation in general will not be able fully differentiate between machine and human translation but might catch some of the more obvious errors from the machine translations. While employees with prior translation experience might be able to catch the machine translation or LLM translation more quickly. Both categories of people might not initially note any distinctions in perception of the target texts beyond surface level errors."

Out of the five final participants in the study, one answered that they had received formal training as a translator and held a degree in translation studies. The other four all answered that they held or were currently in the process of completing a degree in language teaching, particularly English. As figure 14 below shows, the same as figure 7 in the subchapter 4.1.2., all but one study participant answered that they had prior experience with machine translation.

Of those four, three had previously used DeepL, one had used ChatGPT, and one had used another tool, in their case Google Translator. The first part of Hypothesis 2, "Austrian Armed Forces' employees with no prior experience with ChatGPT, DeepL or translation in general will not be able fully differentiate between machine and human translation but might catch some of the more obvious errors from the machine translations", was proven correct in the sense that those participants, who had no prior experience with machine translation in general and those that had never used DeepL or ChatGPT for translation before, struggled during the experiment to make a distinction between DeepL and ChatGPT. All participants, regardless of translation experience, recognised the human translation.

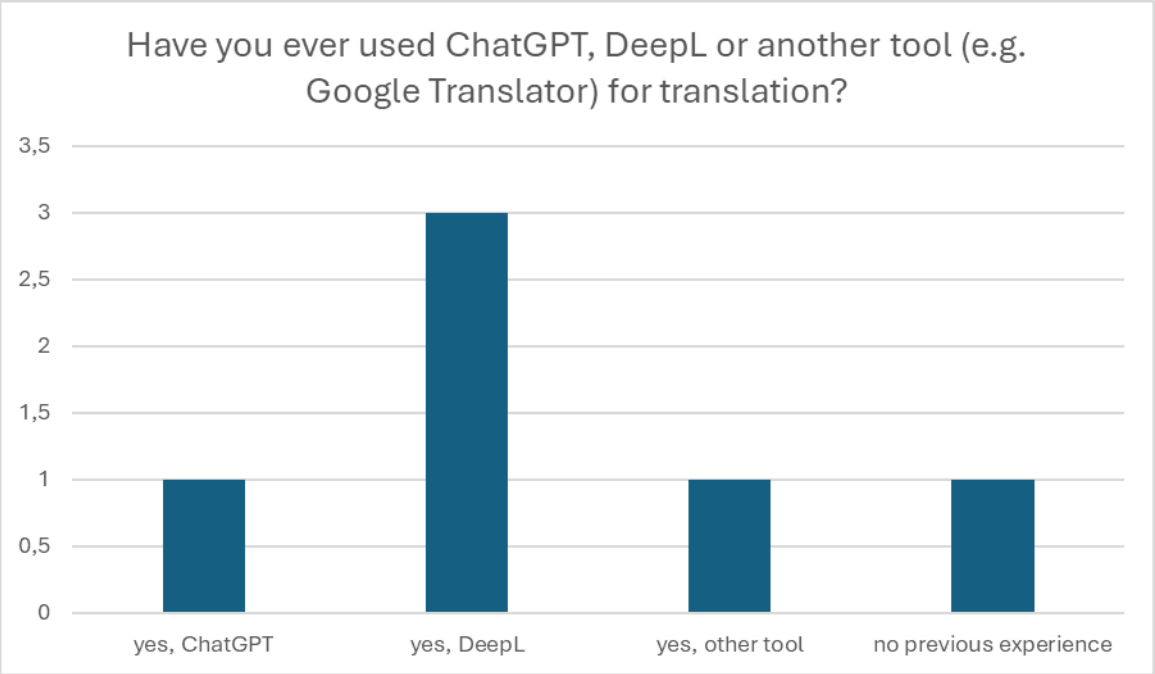

Figure 14: bar chart of participants' answers regarding previous use of machine translation tools

The second part of Hypothesis 2, "[w]hile employees with prior translation experience might be able to catch the machine translation or LLM translation more quickly", also held true for the most part. Of those participants with prior translation experience, all had answered that they had used DeepL before and thus not only knew for what kind of errors to look out for but were also able to successfully differentiate between the translation done by DeepL and the human translation and thus, through the process of elimination, were then able to point out the translation done by ChatGPT.

It seems that having any prior translation experience at all, as well as the level of experience and the familiarity with the domain- and locale-specific terminology and the machine translation or LLM translation tools not only aids translators in recognizing the

origin of a target text, but also aids their confidence when having to guess a text's origin. This also tracks with personal experience and industry standards in relation to training for translators. The more experienced they are, the easier translators – with formal training and degrees in translation or without – might find post-editing and spotting of errors also in human translation in accordance with the two-man rule as laid down in ISO 17100. And the more likely it is, they have strategies in place for dealing with terminology errors or while editing a translation.

Whether people with absolutely no prior translation experience, who have always only been target audience and readers of target texts, would take further research. I would also recommend further studies with a larger participant pool from both a pool of enlisted personnel as well as civilians. It also bears repeating that further studies with different language combinations, perhaps different source languages who then get translated into German, would also be of interest. Particularly lower resource languages would be of interest here, like Albanian and Serbian, due to the Austrian Armed Forces' involvement in United Nations' peace-keeping missions like KFOR.

Interviewee 3 had noted in their interview, and as has been mentioned previously,

> *Und ich finde, gerade wenn du als Mensch übersetzt, denkst du ja eher noch, du denkst da noch anders darüber nach, als wenn du das einfach nur liest. Wenn du das jetzt einfach so auf Deutsch liest, erfasst du halt die Information und denkst dir ja, passt. Aber wenn du es übersetzen musst, dann liest du es und wenn es ein langer Satz ist, dann denkst du nochmal so: okay, was sagt das jetzt wirklich aus? (Interviewee 3: 30)*
>
> *("And I think that especially when you translate as a human being, you tend to think about it differently than when you just read it. If you just read it in German, you simply grasp the information and think to yourself, that's fine. But if you have to translate it, then you read it and if it's a long sentence, then you think again: okay, what does that really say?")*

This difference in cognitive approaches when reading a text purely for information gain as the target audience as opposed to reading a target text as a translator and still being part of the target audience, is also why I want to advocate for further research and studies with a diverse participant pool, as I believe, these cognitive differences would be interesting and could also aid in not only teaching translators how to formulate a target text that is appropriate for the target audience but also to teach non-translators how to spot raw machine translation outputs and perhaps advocate for better end-quality in products.

The final part of Hypothesis 2 stated that "[b]oth categories of people might not initially note any distinctions in perception of the target texts beyond surface level errors." The two categories here are the participants with prior machine translation experience and without prior machine translation experience. As has already been mentioned, there appears to be a correlation between machine translation experience and confidence in differentiating different machine translated texts. As previously noted in their interview, Interviewee 5 had struggled with this differentiation and also with noticing errors beyond the surface terminology level.

As mentioned in subchapter 4.2.2., when asked about the reasons for their rankings, all participants cited terminology – or errors in terminology like with ChatGPT translating "alliance" as "Allianz" instead of "Bündnis" or "service men and women" as "Dienstmänner und -frauen" instead of "Soldatinnen und Soldaten" – as the first clue to the origins of the texts. Two out of five then also cite fluency, audience appropriateness or formatting as reasons for their ranking. However, the prevalence of terminology as reasons for the ranking and as clues for the study participants suggests that a familiarity with the target language's domain-specific terminology can help in assessing both a human and a machine translation. Knowledge of the terminology was not tied to participants' formal teaching and any degrees they held but rather to their experience with translations in general and editing translations in particular. Figure 15 also shows the reasons for the rankings in correspondence with the frequency that they were named by participants.

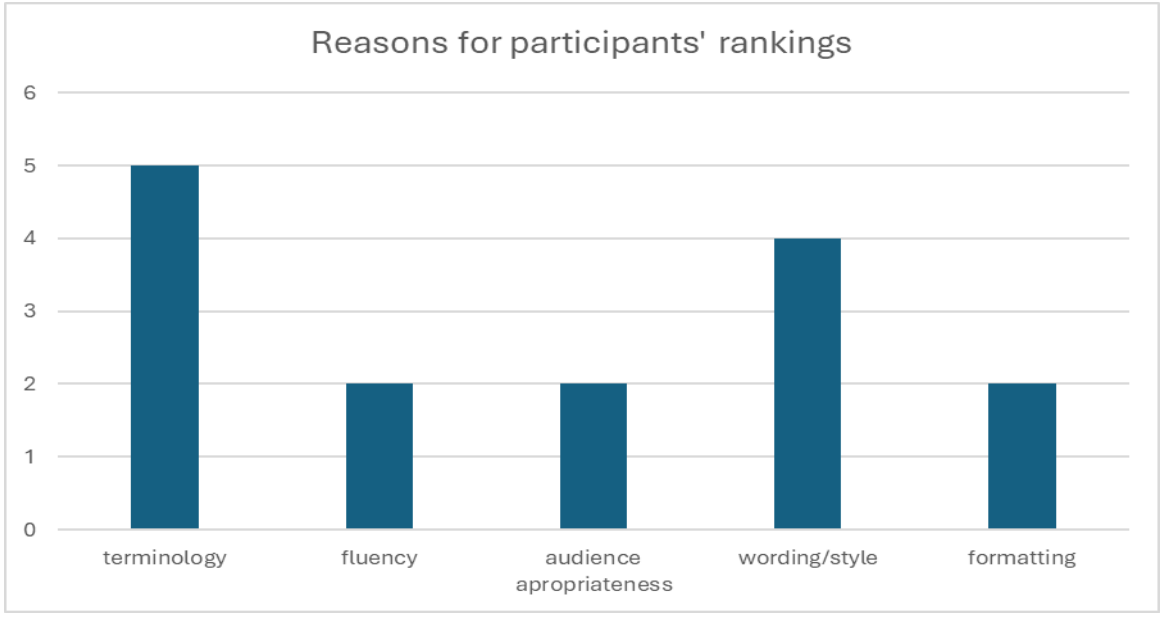

Figure 15: bar chart of categories of reasons for participants' rankings

The last part of Hypothesis 2 was only partially proven correct as Interviewees 3 and 4 had also gone beyond the surface level of the texts and terminology and focussed on syntax and readability of the texts initially. Interviewee 4 also mentioned that

> *Außer dass Mensch sehr stark von DeepL beeinflusst ist offensichtlich und wirklich ganz offensichtlich mit dem Programm gearbeitet hat und dass da auch noch Luft nach oben ist, weil gewisse Dinge man trotzdem nicht sagen würde oder schreiben würde. (Interviewee 4: 45)*
>
> *("Except that the human translation is obviously very strongly influenced by DeepL and the translator has obviously worked with the programme and that there is still room for improvement because you would still not say or write certain things.")*

To Interviewees 3 and 4, there was also still some editing that could be done on the human translated text as well, not just on the machine translations. All participants with prior machine translation experience held the impression that DeepL and Google Translator produced good quality texts that were also appropriate for their purposes. Interviewee 5, for example, had noted that they primarily use Google Translator for gist translation of news broadcasts in languages they did not or not fully understand. Interviewee 3 had also noted a dependency on the goal of the machine translated texts, as previously mentioned in subchapter 4.1.4.

> *Es kommt darauf an, also wenn es nur so allgemeinsprachliche Sachen sind, dann ist es sehr gut, aber alles was dann irgendwie so ins Militärische mehr geht, das kann DeepL halt von der Terminologie her einfach nicht. (Interview 3: par. 10)*
>
> *("It depends, so if it's just general language things, then it's very good, but anything that goes more into military terms, DeepL simply can't do that in terms of terminology.")*

Experience with machine translation and the knowledge of the ultimate goal and purpose of the machine translated text within the target language and for the target audience then plays a part in the impressions the texts make on participants. When asked about their impressions on previous machine translations, the participants with experience answered that they found the quality quite good, especially for gist translations or general-language texts, as Interviewee 3 had also said.

As previously mentioned, experience in both translation in general and machine translation in particular helped participants to make informed choices about the target texts they were presented with. Although I have already written out the disclaimer that this topic not only deserves further research but demands it, I do already think that the results of this thesis and this experiment should help in further training for both translators and target

audience or even language teachers. My research question and hypothesis on this part of the thesis were informed by my own experiences with machine translation and a small experiment with a self-written text in ChatGPT where I was faced with the same issues as the participants in this study.

Ultimately, although this thesis provided an interesting experiment and in my opinion valuable insight into the feasibility of machine translation within the military domain, I do think, the environmental impact of ChatGPT, other Large Language Models and machine translation models like DeepL also needs to be considered. While this thesis was a relatively small experiment and the environmental impact of it perhaps not so great, the training of machine translation models does have a large impact on the environment. And I think, should further research into machine translation within the military domain be conducted – with English as source language and German as the target language or other language combinations in general –, the environmental impact should also be considered and perhaps a pooling of resources between countries should be considered.

5.3. Hypothesis 3 and further remarks from participants

Research Question 3 was “How does knowing a text is a machine translation change their perception of the translation?” The corresponding hypothesis was “After being told of the origins of the target texts, respondents might make changes to their ranking and note differences in their perception of the texts.” Not all participants had final remarks on their rankings and the texts, but the ones that did, brought up interesting points.

Some participants with little experience in the use of ChatGPT had mentioned that they assumed that ChatGPT’s “AI capabilities”, or rather its programming as a Large Language Model, and the bigger data sets for training would have an impact on the translation and ensure a higher quality translation. These responses given by participants solidify the correlation between experience with certain machine translation tools and the ability to recognize said tools, like Interviewee 2, who when told that they had mixed up DeepL and ChatGPT said, “Und dann war auch sozusagen meine Vermutung falsch. Dass ChatGPT, auch Kontextbezogen, die Terminologie rausfinden kann, ist DeepL besser, in diesem Fall klar.” (Interviewee 2: 64) (“And then my guess was also wrong, so to speak. The fact that ChatGPT can find out the terminology, also contextually, DeepL is better, in this case clearly.”) All participants, regardless of their experience with DeepL and ChatGPT translations had

expressed surprise at the high quality of the texts. And almost all participants had said that the errors they found within the translations could be easily fixed with skilful Post-Editors.

Other remarks made by participants included the absence of a consistent strategy to use gender-equitable language, for the human translation as well as the machine translated texts, despite the existence of a regulation for the Ministry of Defence that does not give instructions on how to use gender-equitable language in texts, but does state that one strategy should be chosen and then consistently used throughout the text. While DeepL does not possess the capabilities for the user to enter instructions for a specific strategy of gender-equitable language use, giving those instructions to ChatGPT could be possible but would require a different study.

While Hypothesis 3 stated that participants "might" make chances to their rankings, the fact that those who had mixed up DeepL and ChatGPT did after the reveal of the solution, seems obvious. Interviewee 1, who had assigned the correct labels to the machine translations, later expressed relief that they had guessed correctly, which in turn puts in question the correlation between experience with machine translation tools and the ability to differentiate between different tools. However, with the answers received in this study, it appears more likely that experience with certain machine translation tools like DeepL Translator aids translators, as well as post-editors and evaluators when dealing with machine translated texts.

5.4. Final Conclusion

In conclusion, it can be said, despite the popular assumption that machine translation is better equipped to handle technical texts and terminology, which was also mentioned by translator T in the expert interview, and despite the assumption voiced by some participants in the study that ChatGPT is better at translating than other machine translation tools, that neither of these two tools are particularly well-equipped in handling translations within the military domain from English into German. As the evaluation and the study in this thesis show, terminology remains the stumbling block for both DeepL and ChatGPT. Despite instances where both ChatGPT and DeepL handled the grammar well, there were still some where both tools stuck too closely to the source language syntax. While this is not usually a problem with English and German, in this case, it did make for some awkward phrasing. And, as had already been mentioned, the terminology is still a big obstacle for both tools as it is terminology that is not only specific to the military domain, but also specific to the Austrian Armed Forces and thus differs from the terminology of the German Bundeswehr in some cases. Additionally, the

inability to select the Austrian locale for German within DeepL Translator resulted in some common language terms that would be appropriate for a German target audience but that had bothered the Austrian target audience in the form of the study participants.

Although both DeepL and ChatGPT failed the manual evaluation, responses from study participants were more positive when discussing the quality of the translations. They did find much of the same errors as were found during the manual evaluation but also went deeper into the texts at times and also sometimes criticised the syntax of both machine and human translations. That said, the evaluated piece of text was relatively short with 299 words in English for the source text and it would certainly be interesting for further research to use longer parts of the NATO website text or the whole site on NATO's support for Ukraine. This would be interesting because it would open up the possibility to see whether the quality of the machine translation would decline with increased size of the source text. This is, however, not feasible with only the free versions of ChatGPT and DeepL Translator and would likely require funding for the use of more ChatGPT tokens and the premium version of DeepL Translator. In turn, this would also raise the question on whether the free version of DeepL Translator is different in quality of the raw machine output to the premium version and how.

Due to time and resource constraints, changes had to be made to the original experiment design. Instead of nine to ten participants, there were five in total – excluding translator T as they were tasked with producing the human translation and thus would likely have been biased towards their text and were thus ineligible as a study participant. Instead of interviewing enlisted personnel and civilian employees of the Austrian Armed Forces, study participants were exclusively civilian employees of the Austrian Armed Forces' Language Institute in Vienna. Thus, further research could be done on the perception of enlisted personnel, as well as widening the participant pool in order to gather more data and be able to make more concrete conclusions.

Additionally, if the resources permit, further research could be done on other texts within the military domain, perhaps even more specialised texts within the Austrian Armed Forces, as long as they are not classified, and a reversal of the direction of translation could be done in order to see how well the machine translation tools can handle translating specific texts from the military domain from Austrian German into English. Other possibilities for further research could be the choice of different language combinations altogether or switching the source language from English to another language like Italian, French or

Russian in order to investigate how well the machine translation tools could handle those language combinations within the military domain.

For the further training of translators or post-editors, it should also be said that, regardless of the identity of their employer, they should be aware of any regulations or guidelines towards the use of gender-equitable language in translations and to either translate from scratch accordingly or edit the machine translations to fit the respective guidelines and regulations. While DeepL Translator does not possess the space for the input of specific instructions regarding the use of gender-equitable language in translations, further research could also be done with giving ChatGPT, and perhaps DeepL Write for a comparison of tools, the explicit instructions to use gender-equitable language in a specific way – either by using the gender star or the internal I or to write out both feminine and masculine forms. Even further research could be done on the perception of the target audience when presented with already post-edited machine translation output in comparison to edited and finalised human translations. Due to resource and time constraints, this was not possible for this thesis, although it might have put the machine translation tools at a disadvantage since their raw output was used.

Bibliography

Army in Europe and Africa Publications (2024). *Translations*. https://www.aepubs.eur.army.mil/translations/ (Stand: 20.8.2024).

Asscher, Omri (2023). The position of machine translation in translation studies: A definitional perspective. *Translation Spaces*,*12* (1), 1–20. DOI: 10.1075/ts.22035.ass.

Bundesministerium für Bildung, Wissenschaft und Forschung (2024). *Diplomgrade*. https://www.oesterreich.gv.at/themen/arbeit_beruf_und_pension/titel_und_auszeichnungen/1/Seite.1730507.html (Stand: 2.11.2024).

Center for Intelligent Information Retrieval (2024). *GALE Project | Center for Intelligent Information Retrieval | UMass Amherst*. https://ciir.cs.umass.edu/nightingale (Stand: 21.4.2024).

Council of Europe (2024). *Dealing with propaganda, misinformation and fake news - Democratic Schools for All - www.coe.int*. https://www.coe.int/en/web/campaign-free-to-speak-safe-to-learn/dealing-with-propaganda-misinformation-and-fake-news (Stand: 7.5.2024).

Daems, Joke & Macken, Lieve (2020). Post-Editing Human Translations and Revising Machine Translations. Impact on Efficiency and Quality. In: Koponen, Maarit, Mossop, Brian, Robert, Isabelle S., & Scocchera, Giovanna (Eds.) *Translation Revision and Post-editing: Industry Practices and Cognitive Processes*. London: Routledge. 50–70. https://doi.org/10.4324/9781003096962.

Dijk, Andrea van (2021). *Talk up front: the influence of language matters on international military missions with a particular focus on the cooperation between soldiers and interpreters*. Tilburg: Tilburg University.

DWDS – Digitales Wörterbuch der deutschen Sprache (2016). *Dienstmann – Schreibung, Definition, Bedeutung, Etymologie, Beispiele*. https://www.dwds.de/wb/Dienstmann (Stand: 9.11.2024).

DWDS - Digitales Wörterbuch der deutschen Sprache (2016). *Allianz – Schreibung, Definition, Bedeutung, Etymologie, Synonyme, Beispiele*. https://www.dwds.de/wb/Allianz (Stand: 31.10.2024).

DWDS - Digitales Wörterbuch der deutschen Sprache (2024).a *Bündnis – Schreibung, Definition, Bedeutung, Etymologie, Synonyme, Beispiele*. https://www.dwds.de/wb/B%C3%BCndnis (Stand: 31.10.2024).

DWDS - Digitales Wörterbuch der deutschen Sprache (2024).b *DWDS-Wortverlaufskurve für „Bündnis · Allianz“*. https://www.dwds.de/r/plot/?xrange=1946:2024&window=3&slice=1&q=B%C3%BCndnis&corpus=zeitungenxl (Stand: 31.10.2024).

European Commission (2024). *The EU's Digital Services Act*. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en (Stand: 24.5.2024).

European Court of Auditors (ed.) (2021). Disinformation affecting the EU: tackled but not tamed. https://www.eca.europa.eu/lists/ecadocuments/sr21_09/sr_disinformation_en.pdf.

Garrigoux, Audrey (2022). Disinformation and Russia's war of aggression against Ukraine. https://www.oecd-ilibrary.org/governance/disinformation-and-russia-s-war-of-aggression-against-ukraine_37186bde-en;jsessionid=Ga64suFLZgk1-9OOzL7eju8BaFdtIZVrCKTMVy31.ip-10-240-5-88.

Holland, M.; Schlesiger, C. & Tate, C. (2000). *Evaluating Embedded Machine Translation in Military Field Exercises*. In: White, John S. (Ed.) *Envisioning Machine Translation in*

the Information Future. Berlin, Heidelberg: Springer. 239–247. DOI: 10.1007/3-540-39965-8_27.

ISO (2015). *ISO 17100:2015(en), Translation services — Requirements for translation services*. (Stand: 11.6.2024).

Kenny, Dorothy (2022). Machine translation for everyone: Empowering users in the age of artificial intelligence. Zenodo. DOI: 10.5281/ZENODO.6653406.

Koehn, Philipp (2020). *Neural Machine Translation*. Cambridge University Press. 1st ed. DOI: 10.1017/9781108608480.

Landesverteidigungsakademie des österreichischen Bundesheeres (ed.) (2014). *Militärwörterbuch - Englisch*. Vienna: Bundesministerium für Landesverteidigung und Sport.

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.

Martins Geraldes, Sofia (2020). The Dark Side of Interconnectivity: Social Media as a Cyber-Weapon? In: Moehlecke de Baseggio, Eva, Schneider, Olivia, & Szvircsev Tresch, Tibor (Eds.) *Social Media and the Armed Forces*. Cham: Springer International Publishing. 1st ed. 189–205.

Mathes, Roswitha & Gruber, Nicola (2019). Gleichstellung - Grundlagen im Österreichischen Bundesheer.*14*, 114.

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.

MQM Council (2024). *The MQM Error Typology – MQM (Multidimensional Quality Metrics)*. https://themqm.org/error-types-2/typology/ (Stand: 11.6.2024).

NATO (2024). *Relations with Ukraine*. https://www.nato.int/cps/en/natohq/topics_37750.htm (Stand: 18.4.2024).

Nord, Christiane (2006). *Text Analysis in Translation : Theory, Methodology, and Didactic Application of a Model for Translation-Oriented Text Analysis. Second Edition*. Leiden: Brill. https://uaccess.univie.ac.at/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=160174&site=ehost-live.

Norri-Sederholm, Teija; Norvanto, Elisa; Talvitie-Lamberg, Karoliina & Huhtinen, Aki-Mauri (2020). Misinformation and Disinformation in Social Media as the Pulse of Finnish National Security. In: Moehlecke de Baseggio, Eva, Schneider, Olivia, & Szvircsev Tresch, Tibor (Eds.) *Social Media and the Armed Forces*. Cham: Springer International Publishing. 1st ed. 206–225.

Program Executive Office, Intelligence, Electronic Warfare, and Sensors (PEO IEW&S) (2016). *Machine Foreign Language Translation System*. https://www.army.mil/standto/archive/2016/11/30/ (Stand: 24.5.2024).

Slator (n.d.) The State-of-the-Art in Machine Translation with Language Weaver’s Bart Maczynski. https://open.spotify.com/show/0PJd1KMW6Cxq2IxFX8hfoC (Stand: 30.4.2024).

SYSTRAN by ChapsVision (2024). *SYSTRAN’s history, a pioneer in machine translation technologies*. https://www.systransoft.com/de/systran/ (Stand: 4.4.2024).

Thompson, Brian; Dhaliwal, Mehak Preet; Frisch, Peter; Domhan, Tobias & Federico, Marcello (2024). A Shocking Amount of the Web is Machine Translated: Insights from Multi-Way Parallelism. arXiv. DOI: 10.48550/arXiv.2401.05749.

Yao, Han; Yan, Yusheng; Xiong, Xiaodan; Sun, Mengyang; Dong, Wenxuan; Jiang, Yang; Li, Xingya & Su, Huichao (2020). Machine translation optimization method and system in field of national defense and military industry. https://worldwide.espacenet.com/publicationDetails/biblio?FT=D&date=20201002&DB=EPODOC&locale=&CC=CN&NR=111738022A&KC=A&ND=1.

Appendix

Source Text

“Relations with Ukraine.” Excerpt (NATO 2024)

At the 2016 NATO Summit in Warsaw, the Alliance's measures in support of Ukraine became part of the Comprehensive Assistance Package (CAP), which is designed to support Ukraine's ability to provide for its own security and to implement wide-ranging reforms based on NATO standards, Euro-Atlantic principles and best practices.

Under the CAP, NATO has helped Ukraine transform its security and defence sector for many years, providing strategic-level advice via the NATO Representation to Ukraine and practical support through a range of capacity-building programmes and initiatives. Through these programmes and tailored advice, NATO has significantly strengthened the capacity and resilience of Ukraine's security and defence sector, as well as its ability to counter hybrid threats. NATO and Allies have also provided extensive support to capability development, including through training and education and the provision of equipment.

Complementing the CAP, several **Trust Funds** have been launched since 2014. These Trust Funds provide resources to support capability development and sustainable capacity-building in key areas. Active Trust Fund projects focus on:

- **Command, Control, Communications and Computers (C4)**, which assists Ukraine in reorganising and modernising its C4 structures and capabilities;
- **Medical Rehabilitation**, which seeks to support Ukraine in enhancing its medical rehabilitation system to ensure that long-term sustainable services are provided to patients, including active and discharged Ukrainian servicemen and women and civilian personnel from the defence and security sector; and a
- **Professional Development Programme**, which helps develop the abilities of civilians working in Ukraine's defence and security institutions.

Completed Trust Funds have also supported Ukraine in the areas of military career transition; explosive ordnance disposal (EOD) and countering improvised explosive devices (C-IED); the destruction of small arms and light weapons (SALW), conventional ammunition, and anti-personnel landmines; ammunition stockpile safety management; safe radioactive waste disposal and land restoration; cyber defence; and logistics and standardization.

Target Text A (DeepL Translator 2024)

Auf dem NATO-Gipfel 2016 in Warschau wurden die Maßnahmen des Bündnisses zur Unterstützung der Ukraine Teil des Umfassenden Beistandspakets (Comprehensive Assistance Package, CAP), mit dem die Fähigkeit der Ukraine unterstützt werden soll, für ihre eigene

Sicherheit zu sorgen und weitreichende Reformen auf der Grundlage von NATO-Standards, euro-atlantischen Grundsätzen und bewährten Praktiken durchzuführen.

Im Rahmen der GAP hilft die NATO der Ukraine seit vielen Jahren bei der Umgestaltung ihres Sicherheits- und Verteidigungssektors, indem sie über die NATO-Vertretung in der Ukraine Beratung auf strategischer Ebene und praktische Unterstützung durch eine Reihe von Programmen und Initiativen zum Aufbau von Kapazitäten leistet. Durch diese Programme und maßgeschneiderte Beratung hat die NATO die Kapazitäten und die Widerstandsfähigkeit des ukrainischen Sicherheits- und Verteidigungssektors sowie seine Fähigkeit, hybriden Bedrohungen zu begegnen, erheblich gestärkt. Die NATO und ihre Verbündeten haben auch umfangreiche Unterstützung für die Entwicklung von Fähigkeiten geleistet, u.a. durch Ausbildung und Schulung sowie durch die Bereitstellung von Ausrüstung.

Ergänzend zur GAP sind seit 2014 mehrere Treuhandfonds eingerichtet worden. Diese Treuhandfonds stellen Mittel zur Unterstützung der Entwicklung von Fähigkeiten und des nachhaltigen Aufbaus von Kapazitäten in Schlüsselbereichen bereit. Die aktiven Projekte des Treuhandfonds konzentrieren sich auf:

Kommando, Kontrolle, Kommunikation und Computer (C4), das die Ukraine bei der Reorganisation und Modernisierung ihrer C4-Strukturen und -Fähigkeiten unterstützt;

- Medizinische Rehabilitation, mit der die Ukraine bei der Verbesserung ihres medizinischen Rehabilitationssystems unterstützt werden soll, um sicherzustellen, dass Patienten, einschließlich aktiver und entlassener ukrainischer Soldatinnen und Soldaten sowie zivilem Personal aus dem Verteidigungs- und Sicherheitssektor, langfristige und nachhaltige Leistungen zur Verfügung gestellt werden; und ein

- Berufsentwicklungsprogramm, das die Fähigkeiten von Zivilisten, die in den Verteidigungs- und Sicherheitseinrichtungen der Ukraine arbeiten, weiterentwickelt.

Die abgeschlossenen Treuhandfonds haben die Ukraine auch in den Bereichen militärischer Karrierewechsel, Kampfmittelbeseitigung (EOD) und Bekämpfung improvisierter Sprengsätze (C-IED), Vernichtung von Kleinwaffen und leichten Waffen (SALW), konventioneller Munition und Antipersonenminen, Sicherheitsmanagement von Munitionsbeständen, sichere Entsorgung radioaktiver Abfälle und Bodensanierung, Cyberverteidigung sowie Logistik und Normung unterstützt.

Target Text B (ChatGPT-4 2024)

Auf dem NATO-Gipfel 2016 in Warschau wurden die Maßnahmen der Allianz zur Unterstützung der Ukraine Teil des umfassenden Hilfspakets (Comprehensive Assistance Package, CAP), das darauf abzielt, die Fähigkeit der Ukraine zu stärken, ihre eigene Sicherheit zu gewährleisten und weitreichende Reformen auf Basis von NATO-Standards, euro-atlantischen Prinzipien und bewährten Praktiken umzusetzen.

Im Rahmen des CAP hat die NATO der Ukraine über viele Jahre hinweg bei der Umgestaltung ihres Sicherheits- und Verteidigungssektors geholfen, indem sie strategische

Beratung durch die NATO-Vertretung in der Ukraine und praktische Unterstützung durch eine Vielzahl von Kapazitätsaufbauprogrammen und -initiativen bereitgestellt hat. Durch diese Programme und maßgeschneiderte Beratung hat die NATO die Kapazitäten und die Widerstandsfähigkeit des ukrainischen Sicherheits- und Verteidigungssektors sowie dessen Fähigkeit zur Abwehr hybrider Bedrohungen erheblich gestärkt. NATO und Alliierte haben auch umfassende Unterstützung bei der Fähigkeitsentwicklung geleistet, einschließlich Ausbildung und Schulung sowie Bereitstellung von Ausrüstung.

Ergänzend zum CAP wurden seit 2014 mehrere Treuhandfonds eingerichtet. Diese Treuhandfonds stellen Mittel zur Unterstützung der Fähigkeitsentwicklung und des nachhaltigen Kapazitätsaufbaus in wichtigen Bereichen bereit. Aktive Treuhandfondsprojekte konzentrieren sich auf:

- Führung, Kontrolle, Kommunikation und Computer (C4), das der Ukraine bei der Reorganisation und Modernisierung ihrer C4-Strukturen und -Fähigkeiten hilft;

- Medizinische Rehabilitation, die darauf abzielt, die Ukraine bei der Verbesserung ihres Systems der medizinischen Rehabilitation zu unterstützen, um sicherzustellen, dass langfristig nachhaltige Dienstleistungen für Patienten , einschließlich aktiver und entlassener ukrainischer Dienstmänner und -frauen sowie ziviler Mitarbeiter des Verteidigungs- und Sicherheitssektors, erbracht werden; und ein

- Berufsfortbildungsprogramm, das die Fähigkeiten von Zivilisten, die in den Verteidigungs- und Sicherheitseinrichtungen der Ukraine arbeiten, entwickelt.

Abgeschlossene Treuhandfonds haben die Ukraine auch in den Bereichen militärischer Karriereübergang, Beseitigung von Sprengstoffen (EOD) und Bekämpfung improvisierter Sprengsätze (C-IED), Vernichtung von Kleinwaffen und leichten Waffen (SALW), konventioneller Munition und Antipersonenminen, Verwaltung der Munitionslagerungssicherheit, sichere Entsorgung von radioaktivem Abfall und Landwiederherstellung, Cyberverteidigung sowie Logistik und Standardisierung unterstützt.

Target Text C (Human Translator “T” 2024)

Auf dem NATO-Gipfel 2016 in Warschau wurden die Maßnahmen des Bündnisses zur Unterstützung der Ukraine Teil des Umfassenden Unterstützungsspakets (Comprehensive Assistance Package, CAP), mit dem die Fähigkeit der Ukraine unterstützt werden soll, für ihre eigene Sicherheit zu sorgen und weitreichende Reformen auf der Grundlage von NATO-Standards, euro-atlantischen Grundsätzen und best practices durchzuführen.

Im Rahmen des CAP hilft die NATO der Ukraine seit vielen Jahren bei der Umgestaltung ihres Sicherheits- und Verteidigungssektors, indem sie über die NATO-Vertretung in der Ukraine Beratung auf strategischer Ebene und praktische Unterstützung durch eine Reihe von Programmen und Initiativen zum Kapazitätenaufbau bietet. Durch diese Programme und maßgeschneiderte Beratung hat die NATO die Kapazitäten und die Widerstandsfähigkeit des ukrainischen Sicherheits- und Verteidigungssektors sowie seine Fähigkeit, hybriden Bedrohungen zu begegnen, erheblich gestärkt. Die NATO und ihre Verbündeten haben auch

umfangreiche Unterstützung für die Fähigkeitsentwicklung geleistet, u.a. durch Ausbildung sowie durch die Bereitstellung von Ausrüstung.

Ergänzend zum CAP sind seit 2014 mehrere Treuhandfonds eingerichtet worden. Diese Treuhandfonds stellen Mittel zur Unterstützung der Entwicklung von Fähigkeiten und des nachhaltigen Aufbaus von Kapazitäten in Schlüsselbereichen bereit. Aktive Treuhandfondsprojekte konzentrieren sich auf:

- Führung, Information, Kommunikation, Computersysteme (C4), das die Ukraine bei der Umstrukturierung und Modernisierung ihrer C4-Strukturen und -Fähigkeiten unterstützt;

- Medizinische Rehabilitation, mit der die Ukraine bei der Verbesserung ihres medizinischen Rehabilitationssystems unterstützt werden soll, um sicherzustellen, dass Patienten, einschließlich aktiver und entlassener ukrainischer Soldaten und Soldatinnen sowie ziviles Personal aus dem Verteidigungs- und Sicherheitssektor, langfristige und nachhaltige Leistungen erhalten; und ein

- Programm zur beruflichen Weiterentwicklung, das die Fähigkeiten von Zivilisten, die in den Verteidigungs- und Sicherheitseinrichtungen der Ukraine arbeiten, weiterentwickelt.

Abgeschlossenen Treuhandfonds haben die Ukraine auch in den Bereichen militärischer Karrierewechsel, Kampfmittelbeseitigung (EOD) und Abwehr behelfsmäßiger Spreng- und Brandvorrichtungen (C-IED), Vernichtung von Klein- und Leichtwaffen (SALW), konventioneller Munition und Antipersonenminen, physische Sicherheit und Verwaltung von Munitionslagerbeständen, sichere Entsorgung radioaktiver Abfälle und Bodensanierung, Cyberabwehr sowie Logistik und Standardisierung unterstützt.

Expert Interview, July 18th, 2024

1 **Sophie Grossalber:** Okay. First of all, thank you for agreeing to do the translation and the interview.
2 **T:** No problem. Always a pleasure, never a chore.
3 **Sophie Grossalber:** Just to reiterate, this interview is for my MA thesis on translation perception and all answers will be anonymized if I'm citing them in the thesis.
4 **T:** No problem. No state secrets will be revealed.
5 **Sophie Grossalber:** No, I don't think so. To start, what are your qualifications for translating?
6 **T:** Okay, let's start like this. I did my Matura, my school leaving exam, qualifying for University and so on, or A levels or whatever, in '91. Then, I began my studies of English and German to become a teacher. Now, at that time the system was a lot simpler than it is today. You became a Magister, which now is more or less the equivalent of a Master's degree. And the only difference between studying for the teaching profession and studying for a diploma was one semester. So basically, it was eight months diploma studies and then one semester could be sliced and diced, as you know, added to the various other semesters, where you were taught the basics of didactics, pedagogy, etc. and all this stuff. Which then, of course, meant that at the end of your training, at the end

of your studies, suddenly you were faced with children. And lots and lost of people found out, “It’s not for me”. In the teacher training colleges, which today trains you to become a Bachelor, at that time, this didn’t exist. You had to teach from the second semester. So, a lot of people there, a lot of people attending teacher training colleges, realized very soon that this job wasn’t for them. Grammar school teachers, teachers of high technical, of technical colleges and so on, and commercial colleges, realized this at the end of their studies and suddenly realized they would spend forty years in a profession they didn’t enjoy.

7 I did this. I finished in 1997, and then I immediately, I had my diploma exam. At that time, it was just a diploma exam, not a defensio. On the 26th of September. And on the 29th of September 1997, I started my national service. At that time, eight months. And after eight weeks of basic training, I was transferred here to the National Defence Academy, to the Austrian Armed Forces Language Institute. At that time, it was the language department of the National Defence Academy. And then I was immediately tasked with teaching and testing. The thing is, I wasn’t faced with children. I was faced with grown-ups. And the good thing about grown-ups, if you teach grown-ups, their parents won’t come in and complain. So, that was how I began. And at the beginning it was just teaching and testing. However, when the system then grew, especially because the decision was made that every soldier would need a certain skillset in English. And when people suddenly realized, yes, there’s a world out there that doesn’t necessarily speak German. And bad English might be good enough for communication, but not really good for anything else. We were of course inundated with letters and with texts, etc. Then we had a rather small translation department, and they said, “Look we can’t handle this anymore.” So, I started getting texts to translate. Ninety-nine percent of the time German into English. English into German was, has been, and is very rare. That's how I slowly also began this process of learning how to translate. One good remark was made by a boss of mine who said, “translating isn't a science, it's a skill.” And I think he was very right about this. I've- Nothing wrong with people who study, what was it? Transcultural communication I think is that it?

8 **Sophie Grossalber:** Bachelor. The Master’s is just Translation.

9 **T:** What, whatever, whatever it's called. Very good at awareness raising but as with almost all jobs you need experience, you need to have done things for a number of years. The motivation is wonderful, talent is wonderful, but I once read this wonderful remark, “Hard work will always beat talent if talent doesn't work hard.” An academic approach to things can also be very good, you know? How do you translate, how do you think about text, all these things. But they say you need to have done these things for many, many years, so that you can then also realise, ah, that was bad. But what I did a couple of years ago, I wouldn't do this this way now, etc. And you need somebody, you need a second pair of eyes which is at least as good, if not better than your pair of eyes because this guy, this lady will then tell you: “Mistake, mistake, mistake. That's a problem. I wouldn't do it this way, etc, etc, etc.” And that's how I, why now do testing and teaching and translating. Interpreting sometimes as well which I had to learn for the current minister of defence because she sometimes needs interpretation. But that's not

simultaneous. That's basically just sentence translation, what you would call in German "konsekutiv". Simultaneous translations I cannot do, I haven't had the training, I would never touch this. So that's basically where I come from, so Jack of all trades. Yeah, "needs must", I think, is the phrase.

10 **Sophie Grossalber:** Okay, so do you have any experience with machine translation?

11 **T:** Yes, I remember the first time– It was quite funny, my father is also an English teacher, my mother is an English teacher, so basically we, we could talk about genetic predisposition here. And my father called me a couple of years ago and said, "You know there's a new tool out there, DeepL. That was– the name is very strange, not as tripping off the tongue like Google or whatever. DeepL, I wrote it down and said it's really good." And I said, "Come on, Internet, free of charge, it can't be that good." And he said, "Try it out." And I tried it out and I was really horrified. I–I stared at the computer screen and I saw the end of my career. And for the first time in my life, I thought to myself, "My God, I'm so happy that I'm no longer young." Because the quality – that was a couple of years ago, six or seven – the quality already at that time was, for something that's freely available stupendous. Plus, the translation sounded English. It was not, you know there are certain types of texts where you know, yeah this was not written by a native speaker. It's perfectly correct English but nobody will talk like this. With DeepL this was never the case, and it was really, really, really, really shocking for me. So I, what I then learned was that you then need a really good skill set linguistically in the sense that you need to be able to realise that certain phrases aren't idiomatic but it also requires that you really think and that you don't let the machine pull the wool over your eyes. When you also need to ask yourself is that correct, is that good English or is that, does that just sound nice? In German, "Kann man das so sagen?" So, can you really say that? That was but the thing with DeepL is, is the quality is really that good and, of course, it put the kibosh on generative– what was the generative transformation grammar, whatever Chomsky you know trotted out through the Seventies, Eighties and Nineties. That's all dead. It now works with corpora and artificial intelligence, so yeah. Also the end of linguists. I remember a text I read many, many years ago saying that, what the computer Hal in 2001, the famous science fiction movie, what he says is: "I'm sorry, I'm afraid I can't do that." Whatever the phrase is. That no computer will never be able to say this because you know this, you need certain intelligence. Yeah, that's possible now. So, never say never.

12 **Sophie Grossalber:** Yeah, ChatGPT will tell you when it can't do something.

13 **T:** Exactly. But also for ChatGPT you now need somebody, something… I-I never knew how certain words come back. You need prompters, people who write correct prompts, so that ChatGPT can produce a nice little text. So the prompters only used to be used in theatres but now it's prompters for artificial intelligence.

14 **Sophie Grossalber:** Have you tried any of the other tools like Google Translate?

15 **T:** I once tried Google Translate, it was just, it was abysmally bad. It's very, it's a very funny thing because Google will normally, I mean Google rules the world. I mean, they, they have a problem and within two weeks, I mean, they've solved the problem. My favourite story is that Google picture search was prompted, to come back to your phrase,

by Jennifer Lopez's green dress. Everybody wanted to see Jennifer Lopez's green dress, and people asked, "How can I find this on the internet?" And then Google developed the picture search. The artificial intelligence, the translation programme is way behind. So basically, I'm just waiting for Google to buy DeepL.

16 **Sophie Grossalber:** Yeah, I mean, I think Microsoft has a stake in OpenAI now, so they're probably, they're buying ChatGPT…

17 **T:** Yeah. So I was a bit, I was a bit shocked. Plus what is so funny for me about DeepL is that it relies on the texts collected by linguee. And that was something I only used, to put it sarcastically, to find examples of how not to do something and if anybody had told me they could use these text corpora to produce something of that quality, I would've said, Italian cooking or Italian cuisine: "They make everything out of nothing, and the Brits produced nothing out of everything." So, the raw material isn't always the most important thing.

18 **Sophie Grossalber:** Mhm. As you've translated this text, what do you think are some of the challenges machine translation would face?

19 **T:** I want to congratulate you on your choice of text because it drove me nuts. It reminded me of many, many, many, many, many discussions and it just goes to show that machine translation… The one theory about machine translation is that the machine translation is really, really good at technical texts or at technical language, in the sense that it's a specialised set of terminology. No, it's not. It's getting better there as well but what people who underestimate, especially about military text or texts related to the military sphere, is how much politics plays a role there. Not party politics but questions of how do you refer to something? The military, especially the United States, because of the American tradition of constantly saying, "You can't say that", then followed at lunch by "I can't eat that", of course, also affects the military. So that the lists and lists, you know, of insane military statements. Referring to a Titan II missile as a very large potential disruptive re-entry system is completely insane. Or saying that, you know, in order to liberate the village, it became necessary to destroy it and all the other things. But there are certain things also in German, certain ideas, certain concepts, certain words which every Army, every Navy, all the Armed Forces avoid. It's quite funny, in German, we only have a Defence Academy, in France you have an École de guerre. If anybody had the idea of opening a "Kriegsschule" in Austria, that would be his or her or their very last idea. They would find themselves sitting in the cellar next to the photocopier for the rest of their career. So that's the one thing. The second thing is that one phrase that was always used for when the Western forces left Afghanistan. They didn't talk – nobody said or used the phrase of withdrawing from Afghanistan. They used the phrase of redeployment. I always thought this was wrong because redeployment was a word I normally associate with if you reorganise a company and when you redeploy people in a company. However, it was not totally wrong because first you deploy forces into an area of operations and, so it is logical that you redeploy them home. And the good thing about this word is, the word withdrawal, which is, of course, not a nice thing for armed forces, can be completely avoided.

20 One thing I find obscene with the American armed forces was a term which really horrified me and that was the Mortuary Affairs Collection Point unit. To refer to your own dead soldiers, to your own fallen – but that was, that is the phrase that has always been used, “For the Fallen”, a legendary poem by Laurence Binyon – as Mortuary Affairs, I find obscene. You don't need to pin a medal on every, every chest. A custom which is also something which is very American in the sense that they don't have soldiers, there, they're all heroes. Which is, that begs the question what is somebody who gets the Congressional medal of honour? Is he then a superhero or she? It’s that terminology that for me is insane. But to refer to your own dead soldiers or to any soldiers, dead, as Mortuary Affairs is something I would never do. I find this obscene; I find this offensive. So that was also something where I said no. This is… No. This is something you shouldn’t do. If you start a war, if you go into a military operation you have to expect and accept that people will die. If you don't like this, then don’t start a war or go into an operation. It's a very simple thing. It's terrifying but it's the job of armed forces to fight. Next thing, because in the old days you just said to kill the enemy, now of course, we say engage the enemy. It’s engagement. Which is of course the next big problem because in German the word is “Engagement”. French word. Which… I tell everybody: “Please do not say engagement, especially in a military context. Use commitment.” But the European Union has also reacted a couple of years ago to the word “engagement”, in a European context, is now used, has now been used wrongly for so many years that everybody says, “Germany’s engagement in the Balkans”, or “French engagement in Africa”. They do not refer to any military scenario, or dead people or injured. No, it just means what we refer to as “Engagement”, but they were refer to as commitment. So, you always have to be aware, especially with military texts – what nobody would expect – is that the terminology is very fluid.

21 Words are suddenly very much out of fashion because it no longer reflects the world we want to create linguistically. So, you have to be very flexible, and you have to be constantly aware of changes. Plus, a bit with translations, it's ,the next thing is – like I told you that we no longer have a French École de guerre, again, we have a defence Academy. There was one legendary translation which I always found very, very funny and it was… We – the Germans and Austrians called the “vorderer Rand der Verteidigung”. Basically meant where the enemy, your own forces and the enemy forces would hit. There is a special definition of it, but you don't need to know anything except the word “vorderer Rand der Verteidigung”. This was translated for years as the forward edge of the battle area which is pretty funny because the word battle has nothing to do with the word “Verteidigung”. You see, now it's the forward line of our troops but which is also changed. But it's really, really funny because Germany and Austria for very good historical reasons, of course, always try to avoid the term “war” and “battle” and everything. We were just defending. Which of course went dead when the German minister of defence, the German chancellor said, “Deutschland muss kriegstauglich werden”, certain media people got palpitations and said, “Verteidigungstauglich reicht auch”. It's funny because I don't think the one has such, that is such a big difference

between one and the other. But all of these things are in there, so you have engagement which is a big problem.

22 And then you have a wonderful term in German and that's "Ausbildung". It's beautiful because, of course, if "Ausbildung" has to be translated in English it's either "training" or "education". People would ask me, "What's the difference?" I would say, "your daughter should only have sex education but not sex training". That's fair, that's the easiest thing, everybody remembers this. But, of course, the question is what do soldiers do? In English you always train. You train to become a doctor, you train to become a lawyer, you train to become a soldier, but it was always far more with this practical approach. So, what do you do with soldiers? Do we educate them at the academies, do we train them in the field? It's one of these things and then, of course, "Führung". "Führung" is a wonderful term in German because it covers everything. In English you have to differentiate between leadership and command and control. Leadership is always something – the way we define it as how you interact with people. Do you scream at them, do you motivate them with a machine gun behind them. Or do you motivate them by being good boss or office or whatever. That means that. And then, of course, there's command and control. Now most German speakers have no idea what controlling means. They translate controlling as "kontrollieren" which it is not, it's "nachsteuern". So that's the next thing. This is why we have the "Führungslehrgänge", or used to have them and we always translated this as the command courses. Command course one, command course two, because it deals with staff work. The question is, when an order comes in, was implemented and then you have to check how has this order been implemented, how has the situation changed. That's of course the C2 process. Then in German "Führungsprozesse". Which is a wonderful term because it can cover everything so when we say that "Okay, der kann ja nicht führen". He can't lead or he doesn't understand command. Very, very interesting problem, linguistic problem. And then of course the hierarchy in German: Befehl, Auftrag, Einsatz, Operation. That's beautiful. Befehl is order. Auftrag, mission, and Einsatz? It depends. It depends on whether you are more civilian or military. The United Nations only has missions, the European Union has missions and operations and then we have "Operation" which is, oh, operations again. That can't work because we've already used this. For mission sometimes the term task is used but that's only for special forces etc. That's of a lower levels of at maximum company size. But it's a real headache that you have four beautiful terms in German and a lot of problematic terms in English. And this is why you have to be so careful translating these English texts into German or German into English because you always end up with a big, big FUBAR somewhere in the middle. And, of course, the next problem is that, of course, there's a difference between Austrian military terminology and German military terms, British military terminology and American military. They have different concepts standing behind these terms. This is why going international can create such headaches, and why the worst question you can be asked is, "Wie übersetzt man das?"

23 "How do you translate that?" In which context? Is it military, is it civilian, are there things you would like to avoid saying? People never expect this in a military context.

They expect that military orders are clear, precise, concise, short, to the point. Do it. No. No. Because, of course, as soon as civilians read this – "You can't say that, that's much too aggressive, etc, etc, etc." There's one term, an English term, that has two German translations. That's recoil. Recoil can be the "Rückstoss" of a weapon. And "der Rücklauf" of an artillery gun. When it basically recoils, in German that's "der Rücklauf". But that's the one term in German I know of where we have two translations in German. The rest is always a huge problem from German into English. But this text was very well chosen. I really have to congratulate you on that. It gave me supreme headaches which is what it was designed to do. May I ask why did you choose it?

24 **Sophie Grossalber:** Well, I felt it was very adequate for the current situation with Ukraine. Obviously, we've got security concerns about that too and there also is no official German translation of that page on the NATO website.

25 **T:** Yeah, how do you translate "Sicherheit"?

26 **Sophie Grossalber:** Yeah, security, safety.

27 **T:** Exactly, there's a difference between safety and security. Which is "persönliche Sicherheit". Which is the next headache. So, "Sicherheit" is wonderful in German. Safety and security is a bit of a problem with English. All of these things. So, yeah, this is the things you have to look at. You only realise it if somebody tells you. Or with "Gewalt". Force or violence. Police– our armed forces, our police never use violence. They only use force. But, of course, to get this into the brain of a German speaker is hard, because "Is there a difference?" Yes, there is a very big one because violence will get into gaol for a very long time and it's really… Nobody expects that with military texts. That's quite, quite interesting. They expect… If you say you're going to translate Shakespeare that's when everybody goes "That must be very difficult." Military texts? What's the problem? No, there is one. Especially if you go to the political level, there's political, military, civilian, media level. How do we, how do we refer to something? My favourite story, of course, is it "Ukraine" or "the Ukraine"? In German it's "die Ukraine". I've been told by somebody from Ukraine that they just say "Ukraine". "The Ukraine" would have been part of the USSR which is of course a big no-no. So, forget crossing the T's and dotting the I's. You have to get rid of all of the the's. It's really funny that if you tell this to people, they go "That can't be true". It is. And with the military texts, yes, it is. So, congratulations again.

28 **Sophie Grossalber:** Thank you. I remember when we first talked about the text you said that you noticed, the whole text was basically in British English.

29 **T:** It's NATO, as far as I – my experience showed me that NATO always uses British spelling. So, defence with C etc, etc, etc. Why? So, it doesn't look American. If you tell this to people they will look at you with open eyes and open mouth and say, "You cannot be serious." Yes, I am. It's again a question of perception.

30 **Sophie Grossalber:** Terminology-wise, did you notice any differences in terms from like American or British?

31 **T:** No, no. That's… NATO is not a military alliance, NATO is a political alliance, that's the first thing. If you refer to NATO the military alliance you will get hit over the head with a halibut and a very large one. It's a military alliance, it's, it's a political alliance, not

military. So, that's where it starts and so the terminology, I didn't realise any difference with that. You would also need to ask somebody who has experience working, also the experience of decades of working there. There's one thing that I always enjoyed is the NATO list of acronyms. And that is pretty unbelievable because a normal NATO text or military text is almost unintelligible to a non-NATO person because every second word is an acronym. And so, you spend your entire time reading this text looking up acronyms. If you know the acronyms, if you know what the text is about, then, of course, it's easy to read. But for a non-NATO person or a non-military person it's… You might just as well be offering the text in Hungarian, Finnish or Japanese. It's, it's completely unintelligible. Whether or not this is intended, I don't know. Sometimes it does look that way. The only two acronyms I know is "IVO", "in the vicinity of" and "iot", "in order to". Those are the two things I know the rest is just a blank stare of disbelief.

32 **Sophie Grossalber:** I also noticed you translated some of the terms of the German like the "comprehensive assistance package" or "Führung, Information, Kommunikation, Computersysteme".

33 **T:** Which, which, which nobody gets, which nobody does anymore. It is really if you think that teenagers use every second word, they use an English term for every second word because they're committed to challenging… like "total arg". You haven't seen these texts… it's really… nobody does that. Or management speak. My father was once asked by a student, "Was ist das englische Wort für Marketing?" Which is my favourite question. No, there's no German word for "marketing". So C2, C3, C4, C4 ISR - these are all terms nobody translates this stuff but I don't even think anybody could agree on a translation. It's just, it's just thrown in. Because you asked me to translate this, I, of course, then started translating it. And I ended up finding German home pages of the German Bundeswehr where it said, "Das könnte man so sagen. Ungefähr". There is a German term which is not translated, "innere Führung". That is, that's a German term, fully German. There's also an Academy for this. It has its roots in the events of the Second World War, in the events of the Nazi dictatorship. That… just…. You didn't have a soldier in uniform, but "Bürger in Uniform", so citizens in uniform. There were expectations of officers, of commissioned officers, but also lower ranks that they would say no in the event of an order being given that, to them, was wrong. And that is something as a concept that is very German and so it is always "innere Führung". It is not to be translated; it cannot be. Those are things, as we say in Germany: "Isso. Militärische Begründung mit vier Buchstaben." In English, "Because." "Why is that not translated?" "Because". But, of course, it's… Of course, with these terms just like what's it in marketing, of course. It's now the CFO, the CEO, the COO. "Er ist der CEO dieser Firma." There is no translation. Compliance. Then the other one is the, is the… Sorry, I just forgot it, is that you do everything in your power in line with legal requirements. there's a term for it. Due diligence. "Wir haben die due diligence gemacht." Wunderbar. Same with, „Wir haben die due diligence durchgeführt." Schön für die due diligence. It really is… it's per se insane. But this is… this is a language is created, has been created here where people only use English terms or whatever. Because they think that the other

person knows what they mean and agrees on the definition which most of the time happens, is the case, but sometimes it's not. It's the same with military terms, yeah.

34 **Sophie Grossalber:** Okay, that was it.
35 **T:** That was it?
36 **Sophie Grossalber:** That is it, thank you for your time.
37 **T:** Thank you. Very interesting.

Experiment Interview Script

Teil 1: Übersetzungserfahrung

- Dank + Nachfragen bzgl Einverständnis zur Aufnahme und anonymisierten Zitaten in der Arbeit
- Inwiefern haben Sie als Teil Ihrer Tätigkeit innerhalb des ÖBH mit Übersetzungen von Texten zu tun?
- Haben Sie bereits einmal ChatGPT, DeepL oder ein anderes Tool (z. B. Google Translator) zur Übersetzung genutzt?
- Wenn ja, warum haben Sie sich für eine maschinelle Übersetzung entschieden?
- Wenn ja, was war Ihr Eindruck von der Übersetzung?
- Wenn nein, warum haben Sie sich gegen eine maschinelle Übersetzung entschieden?

Teil 2: Ranking

Kommen wir nun zum eigentlichen Teil des Interviews. Ich werde Ihnen gleich drei Übersetzungen in Teilen vorlegen. Eine davon wurde mit ChatGPT angefertigt, eine mit DeepL und eine von einem professionellen Übersetzer. Ich werde Ihnen nicht sagen, welche Übersetzung aus welcher Quelle stammt. Die Übersetzungen sind alle derselbe Auszug der NATO-Webseite über das Comprehensive Assistance Package für die Ukraine.

- Reihen Sie die Textschnipsel bitte von links nach rechts auf. Links ist für Sie „am ehesten menschliche Übersetzung", rechts „am ehesten maschinelle Übersetzung".
- Warum haben Sie sich für dieses Ranking entschieden?
- Was fällt Ihnen an den Texten auf?
- Haben Sie sonst noch Anmerkungen zu den Übersetzungen?

Teil 3: Auflösung und Follow-up

Kommen wir nun zur Auflösung. Bei den Textteilen von Text A handelt es sich um die Übersetzung von DeepL, Text B wurde mithilfe von ChatGPT erstellt und Text C ist die Humanübersetzung.

- Inwiefern deckt sich die Auflösung mit Ihrem Ranking?
- Haben Sie Anmerkungen zu Ihrem Ranking?
- Würden Sie mit dem Wissen jetzt Änderungen an Ihrem Ranking vornehmen?
- Wenn ja, welche?
- Wenn nein, warum nicht?
- Beenden des Interviews

Experiment Interview 1, September 16th, 2024

Sophie Grossalber: So, jetzt erstmal danke gerne fürs Bereiterklären fürs Interview.
Interviewee 1: Gerne.
Sophie Grossalber: Und. Nur noch mal für die Aufnahme, das ist eh alles ok, dass es aufgenommen wird und transkribiert und anonymisiert in der Arbeit zitiert.
Interviewee 1: Ja.
Sophie Grossalber: Passt.
Sophie Grossalber: Dann fangen wir gleich einmal an. Inwiefern haben Sie als Teil ihrer Tätigkeit innerhalb des österreichischen Bundesheeres mit Übersetzungen von Texten zu tun?
Interviewee 1: Ich habe nicht so oft mit direkten Übersetzungen zu tun, weil ich keine Übersetzerin bin. Ich bin eigentlich Englischlehrerin, aber ab und zu habe ich doch was damit zu tun, wenn ich was übersetzen muss, so kleine Sachen hin und wieder für Leute, oder ...
Interviewee 1: es haben Leute schon was übersetzt und ich gehe dann drüber und schaue ob, das ...
Interviewee 1: So sinnvoll ist oder Sinn macht. Oder was mir auch ein, zwei Mal untergekommen ist, ist, dass jemand etwas
Interviewee 1: mit DeepL übersetzt hat und ich dann drüber geschaut habe, ob das Sinn macht, das gut übersetzt ist. Sowas, aber meine Haupttätigkeit ist eigentlich nicht übersetzen.
Sophie Grossalber: Haben Sie selbst bereits einmal ChatGPT, DeepL oder ein anderes Übersetzungstool genutzt?
Interviewee 1: Ja, ich
Interviewee 1: Habe DeepL schon sehr oft benutzt.
Interviewee 1: Ich verwende es meistens zum... Ja, gebe ich meistens den Text ein und dann schaue ich, ob das, also schaue ich noch mal den Originaltext und die Übersetzung an und schaue, dass es so Sinn macht, ob das, ob ich das anders vielleicht sagen würde, ob das vielleicht den Sinn verzehrt, solches.
Interviewee 1: Mit ChatGPT habe ich glaube ich noch nichts.... ChatGPT habe ich noch nicht zum Übersetzen benutzt. Nein.
Sophie Grossalber: Mhm.

17 **Sophie Grossalber:** Warum haben Sie sich für eine maschinelle Übersetzung entschieden? Für die Texte?
18 **Interviewee 1:** Weil...
19 **Interviewee 1:** Ja, es geht schnell und es sind meistens Ideen dabei,
20 **Interviewee 1:** Auf die ich selbst nicht kommen würde. Also manchmal tue ich auch Sachen selbst übersetzen und dann.
21 **Interviewee 1:** Fällt mir eine gewisse Formulierung nicht ein, wie man das am besten auf Englisch sagt? Mir fällt zwar etwas ein, aber ich weiß genau, es ist nicht die perfekte Übersetzung, es gäbe wahrscheinlich einen besseren Ausdruck und dann verwende ich.
22 **Interviewee 1:** DeepL weil da sind meistens mehrere Vorschläge und ich kann mir dann, ich kann mir anschauen, was für Vorschläge kommen und oft ist dann was dabei was.
23 **Interviewee 1:** mir nicht gleich eingefallen wär.
24 **Sophie Grossalber:** Also was würden Sie sagen, war Ihr Eindruck von den Übersetzungen?
25 **Interviewee 1:** Die.
26 **Interviewee 1:** Deep L.
27 **Interviewee 1:** Die sind eigentlich meistens ziemlich gut. Ich war eigentlich oft überrascht, wie gut das eigentlich funktioniert, jedenfalls deutsch, Englisch und englisch, deutsch, wie gut das funktioniert.
28 **Interviewee 1:** Manchmal sind Sachen dabei.
29 **Interviewee 1:** Bestimmte.
30 **Interviewee 1:** Die glaube ich, eine Maschine nicht wissen kann. Also weiß ich nicht, wenn zum Beispiel.
31 **Interviewee 1:** Ein kulturelles Wissen dazu kommt, dann weiß das DeepL nicht. Solche Sachen sind dann meistens nicht perfekt oder manchmal kann man auch vom Originaltext nicht wirklich schließen. Also wenn man jetzt nur einen Satz zum Beispiel eingibt, kann man nicht schließen, was genau jetzt gemeint ist, wo es mehrere Möglichkeiten gäbe, wo man dann ergänzen muss, das Wissen, was wirklich jetzt gemeint ist. Also wenn zum Beispiel Insiderwissen oder sowas gefragt ist.
32 **Sophie Grossalber:** Kommen wir nun zum eigentlichen Teil des Interviews. Ich werde Ihnen gleich drei Übersetzungen in Teilen vorlegen und eine davon wurde mit ChatGPT angefertigt, eine mit DeepL und eine von einem professionellen Übersetzer. Ich sage Ihnen aber nicht welche welche ist.
33 **Sophie Grossalber:** Die Übersetzungen sind alle derselbe Auszug von der NATO Website über das Comprehensive Assistance Package für die Ukraine.
34 **Sophie Grossalber:** Und ich würd Sie bitten, dass Sie die Texte dann von links nach rechts aufreihen oder auf den Zettel schreiben, welche für Sie am ehesten menschliche Übersetzung und am ehesten maschinell ist.
35 **Interviewee 1:** Ist alles klar, die Originalsprache ist Englisch, ok.
36 **Sophie Grossalber:** Ist Englisch.
37 **Interviewee 1:** Danke.
38 **Interviewee 1:** Das ist noch nicht meine endgültige Aufreihung. Das ist zum Anschauen.

39 **Interviewee 1:** Also eine wurde von DeepL, eine ChatGPT und eine von einem menschlichen Übersetzer gemacht, okay, und ich reihe sie auf nach also ich.
40 **Interviewee 1:** Schreibe dann einfach dazu oder sag welche was von wem ist. Ok.
41 **Sophie Grossalber:** Mhm.
42 **Interviewee 1:** Die sind.
43 **Interviewee 1:** Da ist nur der Text eingegeben worden in DeepL und es wurden keine nach, also nichts mehr nachverändert? Okay.
44 **Sophie Grossalber:** Genau.
45 **Sophie Grossalber:** Die rohe maschinelle Übersetzung.
46 **Interviewee 1:** Darf ich das auf dem Zettel markieren oder aufschreiben oder ist das eher nicht gedacht?
47 **Sophie Grossalber:** Vielleicht eher auf dem Collegeblock.
48 **Interviewee 1:** Okay, passt.
49 **Interviewee 1:** Wie lange habe ich dafür Zeit? Ist das eingeschränkt?
50 **Sophie Grossalber:** So viel Sie brauchen.
51 **Interviewee 1:** Ich habe jetzt.
52 **Interviewee 1:** Ein Urteil gefällt.
53 **Interviewee 1:** Ich glaube, dass der letzte Text, der Text C von einem professionellen Übersetzer oder einer professionellen Übersetzerin geschrieben wurde.
54 **Interviewee 1:** Und ich glaube, dass Text A und Text B maschinell übersetzt worden sind.
55 **Interviewee 1:** Ich weiß leider… Ich hab noch nicht so viel von ChatGPT gesehen. Ich nehme an, dass Text A von DeepL übersetzt worden ist und Text B von ChatGPT.
56 **Sophie Grossalber:** Warum?
57 **Interviewee 1:** Ich glaube, dass Text C … hat für mich passendere Formulierungen.
58 **Interviewee 1:** Also die…
59 **Interviewee 1:** Zum Beispiel ist im Text B, im zweiten Absatz steht NATO und Alliierte.
60 **Interviewee 1:** Im Text C steht die NATO und ihre Verbündeten.
61 **Interviewee 1:** Was logischer ist, weil da geht es ja nicht um die Alliierten, sondern um Verbündete von der NATO.
62 **Interviewee 1:** Es geht ja nicht um den Zweiten Weltkrieg, deswegen… Das kann Text B nicht wissen, glaube ich, vielleicht aus dem Kontext.
63 **Sophie Grossalber:** Mhm.
64 **Interviewee 1:** Ja, ich weiß nicht, ob GPT vielleicht da besser ist. Vielleicht weil da steht, im Jahr 2016 am Anfang und dann später 2014 ob das vielleicht hätte, darauf schließen können, ich weiß nicht, aber eben weil es keine Alliierten sind, habe ich darauf schließen können, dass der Text wahrscheinlich maschinell übersetzt worden ist. Im Text A steht auch NATO und ihre Verbündeten, die NATO und ihre Verbündeten.
65 **Interviewee 1:** Genau also das heißt, das ist wieder anders übersetzt. Aber ich habe eben ja für mich, ich weiß, dass DeepL ziemlich gut ist. Deswegen habe ich, glaube ich, ist der Text A von DeepL, weil es war dann zum Beispiel auch…
66 **Interviewee 1:** Im vierten Absatz, da geht es um Kommando, Kontrolle, Kommunikation und Computer.

67 **Sophie Grossalber:** Ja.
68 **Interviewee 1:** Ja, Command and Control ist die Frage, wie man das übersetzt.
69 **Interviewee 1:** Es wurde im Text C als Führung übersetzt, was ich, glaube ich, passender finde als Kommando.
70 **Interviewee 1:** Im Text B steht auch Führung.
71 **Interviewee 1:** Führung, Kontrolle, Kommunikation. Im Text C steht Führung, Information, Kommunikation, das finde ich ist…
72 **Interviewee 1:** Mit dem Wissen übersetzt, was wahrscheinlich… Wo eine Maschine nicht unterscheiden kann.
73 **Interviewee 1:** Deswegen glaube ich, dass Text C nicht maschinell übersetzt worden ist.
74 **Interviewee 1:** Und dann im Text B steht Dienstmänner und -frauen im eins, zwei, drei, vier, fünften Absatz.
75 **Interviewee 1:** Was eigentlich nicht gesagt wird.
76 **Interviewee 1:** Und im Text A steht Personal, glaube ich. Moment.
77 **Interviewee 1:** Zivilem… Sowie zivilem Personal.
78 **Interviewee 1:** Ziviles Personal steht auch im Text C. Das ist für mich passender.
79 **Interviewee 1:** Ja.
80 **Interviewee 1:** Und dann noch später steht was von… In Text A und Text B im letzten Absatz von improvisierten Sprengsätzen.
81 **Interviewee 1:** Das klingt für mich ein bisschen fremd.
82 **Interviewee 1:** Wahrscheinlich ist es auf Englisch „improvised“ oder sowas.
83 **Interviewee 1:** Im Text C steht.
84 **Interviewee 1:** Abwehr behelfsmäßiger Spreng- und Brandvorrichtungen, was finde ich…
85 **Interviewee 1:** Das ein bisschen passender trifft, vielleicht, weil „improvised“ auf Englisch vielleicht dann ein bisschen eine andere Bedeutung hat, aber das macht für mich mehr Sinn.
86 **Interviewee 1:** Die Formulierung Klein- und Leichtwaffen.
87 **Interviewee 1:** Finde ich stimmiger als in den anderen Texten. Klein, also in B steht Kleinwaffen und leichten Waffen und in Text A steht auch Kleinwaffen und leichten Waffen.
88 **Interviewee 1:** Glaube Klein- und Leichtwaffen ist einfach, klingt, klingt einfacher und […] stimmiger, ja, deswegen also ich bin mir nicht sicher.
89 **Interviewee 1:** Für mich ist Text C nicht maschinell übersetzt und die anderen zwei maschinell übersetzt, obwohl.
90 **Interviewee 1:** Ja, das ist auch die Frage.
91 **Interviewee 1:** Ob welches jetzt […] Ich weiß eben nicht wie gut ChatGPT übersetzt, aber weil ich DeepL kenne, das eigentlich ganz gut ist, glaube ich, dass Text A DeepL ist und B von ChatGPT. Aber sicher bin ich mir auch nicht, aber das ist was ich jetzt aus dem schließe.
92 **Sophie Grossalber:** Ja, Sie hatten Recht. Text A ist DeepL, B ChatGPT und C ist die Humanübersetzung.

93 **Sophie Grossalber**: Haben Sie sonst noch irgendwelche Anmerkungen zu Ihrem Ranking?
94 **Interviewee 1:** Na, ich finde es eigentlich cool, dass ich das so herausgelesen hab.
95 **Interviewee 1:** Nein, also was mir auch aufgefallen ist, ist, dass in den maschinellen Übersetzungen manchmal zwei Leerzeichen waren, glaube ich. Das ist jetzt was Kleines und ich glaub.
96 **Interviewee 1:** Von den Menschen übersetzten sind eigentlich immer ist es eigentlich immer nur ein Leerzeichen, aber.
97 **Interviewee 1:** Vielleicht hat es…
98 **Sophie Grossalber:** Das kann auch vom Kopieren sein, ja.
99 Interviewee 1: Wenn ich das so, nein, aber sonst denke ich, bin ich froh, dass ich richtig geraten hab.
100 **Interviewee 1:** Ja.
101 **Interviewee 1:** Mhm.
102 **Sophie Grossalber:** Ja, noch zu der Anmerkung mit den Dienstmännern und Frauen.
103 **Sophie Grossalber:** Ich habe das nachgeschlagen und das wäre eher so.
104 **Sophie Grossalber:** Lehensmann also eher im Mittelalter benutzt worden im deutschsprachigen Raum und.
105 **Sophie Grossalber:** Heutzutage nutzt man es nicht, sondern einfach Personal oder Soldatinnen und Soldaten.
106 **Interviewee 1:** Ja, vielleicht was. Was vielleicht auch, weil es mir jetzt einfällt ist, ich hab geschaut ob.
107 **Interviewee 1:** Mhm.
108 **Interviewee 1:** Wegen dem gendergerechten Sprachgebrauch.
109 **Interviewee 1:** Ob da vielleicht irgendwelche Clues, irgendwelche Hinweise darauf sind, was von wem übersetzt worden sind. Aber ich glaube, teilweise wurde gegendert und teilweise nicht. Also da steht zum Beispiel in Text A.
110 **Interviewee 1:** Aktiver und entlassener ukrainischer Soldatinnen und Soldaten.
111 **Interviewee 1:** Genau. Im Text B, Dienstmänner und Frauen. Und im Text C auch wieder entlassener ukrainischer Soldaten und Soldatinnen. Und dann steht aber vorher wieder in allen Texten, glaube ich, Patienten.
112 **Interviewee 1:** Also ich weiß nicht nach welche, also wie das übersetzt wird.
113 **Interviewee 1:** Ja, interessant. Also ich glaube im Text B steht dann auch Patienten, warum dann Soldatinnen und Soldaten übersetzt wird, aber nicht Patientinnen und Patienten, aber es hat also das hat … von dem aus habe ich nicht schließen können.
114 **Sophie Grossalber:** Mhm.
115 **Interviewee 1:** Das ist interessant.
116 **Sophie Grossalber:** Ja.
117 **Sophie Grossalber:** Da habe ich mir auch den Leitfaden zu Gendergerechter Sprache oder so vom Bundesministerium angeschaut. Und sie sagen auch, also durchgängig gendern oder am Anfang sagen, es wird die männliche Form benutzt und gilt aber für alle.

118 **Interviewee 1:** OK, ja, deswegen wahrscheinlich ja. Ist interessant, dass sie sich in dem Punkt, die drei, die Maschine und der Mensch einig sind.
119 **Interviewee 1:** Ja, aber sonst fällt mir jetzt nichts mehr ein.
120 **Sophie Grossalber:** Gut. Danke. Dann beenden wir das Interview an dieser Stelle.
121 **Interviewee 1:** Auch danke.

Experiment Interview 2, October 2nd, 2024

1 **Sophie Grossalber:** Gut.
2 **Sophie Grossalber:** Also erstens einmal danke fürs Bereiterklären fürs Interview.
3 **Interviewee 2:** Bitte gerne.
4 **Sophie Grossalber:** Und nur noch einmal.
5 **Sophie Grossalber:** for the record...
6 **Sophie Grossalber:** Sie sind einverstanden mit der Aufnahme und dem anonymisierten Zitieren des Interviews in der Arbeit.
7 **Interviewee 2:** Jawohl.
8 **Sophie Grossalber:** Perfekt. Dann fangen wir gleich mal an. Inwiefern haben Sie als Teil ihrer Tätigkeit innerhalb des österreichischen Bundesheeres mit Übersetzungen zu tun?
9 **Interviewee 2:** Insofern, dass ich sie koordiniere und...
10 **Interviewee 2:** Die entsprechenden Übersetzungsanträge den Personen zuteile, die verfügbar sind und in der Lage sind, den Auftrag durchzuführen.
11 **Sophie Grossalber:** Also in dieser Tätigkeit haben Sie schon einmal maschinelle Translation benutzt?
12 **Sophie Grossalber:** Zum Beispiel fürs Analysieren von Texten?
13 **Interviewee 2:** Nein, nein.
14 **Interviewee 2:** Es kann sein, dass das früher passiert ist.
15 **Interviewee 2:** Aber nicht für die Standard Übersetzungen, die wir hier machen.
16 **Interviewee 2:** Es kann in anderen Settings vorkommen, dass man zur Textinhaltserschließung grobe maschinelle Übersetzung darüber laufen lässt, wenn es nicht geheim ist. Aber das ist nicht unser Anwendungsgebiet.
17 **Sophie Grossalber:** Dann gehen wir gleich zum zweiten Teil.
18 **Sophie Grossalber:** Ich habe Ihnen einen Zettel hingelegt, und einen Stift.
19 **Sophie Grossalber:** Zum Notizen machen.
20 **Sophie Grossalber:** Ich werde Ihnen gleich drei Übersetzungen in Teilen vorlegen, eine davon wurde mit ChatGPT angefertigt, eine mit DeepL und eine von einem professionellen Übersetzer. Ich werde nun allerdings nicht sagen, welche Übersetzung aus welcher Quelle stammt und würde Sie bitten, die ....
21 **Interviewee 2:** Mhm.
22 Sophie Grossalber: Texte entsprechend zu reihen, was Sie denken. Was ist ChatGPT, DeepL und menschliche Übersetzung?
23 **Interviewee 2:** In dieser Reihenfolge.
24 **Sophie Grossalber:** Es ist egal. Also ich habe die Texte markiert mit Text A, B und C.
25 **Sophie Grossalber:** Und dann werde ich nachher noch.

Sophie Grossalber: Paar Fragen stellen.
Sophie Grossalber: Der Text ist ein Ausschnitt von der NATO Website zum Comprehensive Assistance Package für die Ukraine.
Interviewee 2: Der Quelltext ist immer derselbe.
Sophie Grossalber: Genau, der ist immer Englisch.
Interviewee 2: Ein englischer Text, ja, ok, immer derselbe Quelltext auf Englisch.
Sophie Grossalber: Ja.
Sophie Grossalber: Normalerweise schreiben die Leute dann auf dem Zettel auf und dann gehen wir alles der Reihe nach.
Interviewee 2: Interessant.
Interviewee 2: Sehr interessant.
Interviewee 2: Darf ich noch mal wissen, was es bedeutet, ChatGPT?
Interviewee 2: DeepL und ChatGPT, achso, ChatGPT. Und die Humanübersetzung?
Sophie Grossalber: Von einem professionellen Übersetzer.
Interviewee 2: Ein fachkundiger Übersetzer, oder?
Interviewee 2: Ich soll jetzt einen Tipp abgeben.
Interviewee 2: Da wäre es hilfreich, etwas mehr zu wissen über.
Interviewee 2: die Fähigkeiten von ChatGPT.
Interviewee 2: Ich habe mit ChatGPT noch zu wenig gearbeitet, aber…
Interviewee 2: ChatGPT macht Übersetzungen.
Interviewee 2: Das wusste ich nicht.
Interviewee 2: Daher ist es schwierig jetzt für mich.
Interviewee 2: Ob ChatGPT oder DeepL über mehr…
Interviewee 2: Oder besser in der Lage ist, die richtige Technologie zu finden. Aber nachdem ich sage mal…
Interviewee 2: Ich nehme mal an, dass ChatGPT über mehr AI-Kapazität verfügt, Dinge herauszufinden ist es vielleicht…
Interviewee 2: Kann vielleicht ChatGPT besser die richtigen Begriffe zuordnen? Also mein Tipp lautet, dass die…
Interviewee 2: Text C die Human Übersetzung ist, weil…
Interviewee 2: Weil die ist flüssig, lesbar und weil ich davon ausgehe, dass die Person…
Interviewee 2: Auch recherchieren kann und nicht einfach irgendwie…
Interviewee 2: Also auch tatsächlich die richtigen Quellen verwendet, um zu recherchieren. Dann würde ich sagen, dass das eine sehr gut recherchierte Übersetzung ist mit den Begriffen, zum Beispiel „Behelfsmäßige Spreng- und Brandvorrichtungen“, ist eine typisch österreichische…
Interviewee 2: Also ein typisch österreichisches Wording, das man wahrscheinlich nur verwenden kann, wenn man, wenn man die österreichischen Vorschriften kennt, beziehungsweise Fachliteratur, also österreichischer Provenienz verfügt.
Interviewee 2: B und A haben teilweise ein bisschen…
Interviewee 2: Unglückliche Benennungen. Zum Beispiel, der Text B hat den Begriff „Dienstmann“, der im Original ziemlich sicher mit „service men“ oder „service

personnel“ bezeichnet ist und „Dienstmann“ gibt es nicht. Ja, daher ist das mein Indikator dafür, dass das ein…

57 **Interviewee 2:** Doch nicht so smartes Machine Translation Tool ist, das den Text B produziert hat und deswegen war jetzt meine Frage: Ich weiß es nicht ob DeepL oder ChatGPT besser in der Lage ist abzugreifen, dass natürlich „service personnel“

58 **Interviewee 2:** Nicht „Dienstmänner“ sind. Also, nachdem ich das nicht genau weiß, war das ein bisschen ein Unsicherheitsfaktor für mich, aber ich vermute trotzdem, dass B DeepL ist und A ChatGPT.

59 **Interviewee 2:** Bei ChatGPT vermute ich, ist es noch besser in der … also noch weiterentwickelt. Die AI, die in der Lage ist, vielleicht herauszufinden,

60 **Interviewee 2**: Was ist „service men“, was ist „service men“ nicht, also, dass es hier um Soldatinnen und Soldaten geht? Ja, das ist meine Begründung des Tipps.

61 **Sophie Grossalber:** Okay, kommen wir zur Auflösung.

62 **Sophie Grossalber:** Text A ist tatsächlich DeepL, Text B ist ChatGPT und Text C die Humanübersetzung.

63 **Interviewee 2:** Dann waren diese beiden hier dem geschuldet, dass ich ganz einfach mit ChatGPT noch nie gearbeitet habe, was Übersetzungen angeht. Ich habe immer nur Dinge gefragt, ja, weiß ich nicht, um es halt auszuprobieren, als Gag, ja, aber nie tatsächlich noch verwendet für solche Zwecke.

64 **Interviewee 2:** Und dann war auch sozusagen meine Vermutung falsch. Dass ChatGPT, auch Kontextbezogen, die Terminologie rausfinden kann, ist DeepL besser, in diesem Fall klar.

65 **Sophie Grossalber:** Ich glaube, es liegt auch daran, dass ChatGPT nicht explizit dafür trainiert wurde.

66 **Sophie Grossalber:** Zu übersetzen. DeepL wurde ja explizit dafür programmiert und hat auch die Trainingsdaten und alles und bei ChatGPT ist es einfach quasi so nebenher zugekommen.

67 **Interviewee 2:** Ja, das kann dann sein.

68 **Interviewee 2:** Ja, hätte ich mich damit näher auseinandergesetzt vorher, hätte ich das wahrscheinlich dann auch so beurteilt. Ja, ja.

69 **Sophie Grossalber:** Mhm.

70 **Sophie Grossalber:** Ja, haben Sie sonst noch irgendwelche Anmerkungen zu Ihrem Ranking? Oder sind Ihnen irgendwelche Dinge aufgefallen?

71 **Interviewee 2:** Zu den Texten. Naja, es ist…

72 **Interviewee 2:** Grundsätzlich bin ich überrascht, wie die hohe Qualität, ja, wenn die beiden anderen Texte, also jetzt A und B, wenn es tatsächlich ohne Post editing gemacht wurde, da bin ich sehr überrascht von der hohen Qualität. Es ist, wenn es der rohe Output ist… Puh, Chapeau, dann kann man das wirklich machen mit einem hochqualitativen, also mit einer Person, die…

73 **Interviewee 2:** Die sehr gut ist.

74 **Interviewee 2:** In Fachsprache.

75 **Interviewee 2:** Natürlich auch über die Textkompetenz verfügt einen Text noch vielleicht zu glätten hier und da.

76 **Interviewee 2:** Wenn man, wenn man das mit Postediting macht, dann ist das, dann ist das ja wirklich faszinierend, wie hoch die Qualität ist.
77 **Interviewee 2:** Aber die Humanübersetzung ist trotzdem besser. Und wie gesagt, wenn man jetzt über A und B.
78 **Interviewee 2:** Wenn man den Übersetzer drüber gehen lasst, der C produziert hat, dann ist das genauso perfekt. Natürlich ja, und weil es ist natürlich eine Zeitersparnis.
79 **Interviewee 2:** Natürlich ist es nicht möglich, im öffentlichen Dienst einfach so.
80 **Interviewee 2:** So ein Tool zu verwenden.
81 **Interviewee 2:** Aus Sicherheitsgründen.
82 **Interviewee 2:** Aber ja, überraschend wirklich sehr interessante Arbeiten zusammen.
83 **Sophie Grossalber:** Ich fand es auch sehr spannend, dass es eigentlich von der NATO-Webseite, jetzt, unabhängig von der Unterseite zum Comprehensive Assistance Package. Es gibt keine offizielle deutsche Übersetzung, was ich auch sehr spannend finde, weil, gut, Deutschland ist NATO-Mitglied, Österreich ist Teil der PfP.
84 **Sophie Grossalber:** Ist die Frage, warum? können wir uns nicht einigen auf Terminologie, oder?
85 **Interviewee 2:** Bei der Englischen, weil Englisch Arbeitssprache ist.
86 **Interviewee 2:** Das ist, das ist ein Text, der zur Kommunikation, wenn ich das, ich weiß nicht, wie heißt dieser Text?
87 **Sophie Grossalber:** Das ist die Infoseite von der NATO zur Situation in der Ukraine.
88 **Interviewee 2:** Ja, also das ist eine Seite zur Kommunikation mit der Öffentlichkeit.
89 **Interviewee 2:** Und somit.
90 **Interviewee 2:** Wäre vielleicht, wenn die NATO in ihrer Policy drin hätte, die deutschsprachige Bevölkerung ihrer Mitgliedsländer gezielt zu informieren.
91 **Interviewee 2:** Dann würden sie es vielleicht übersetzen, aber an sich, glaube ich, wird das grundsätzlich nicht gemacht.
92 **Sophie Grossalber:** Mhm.
93 **Interviewee 2:** Was anderes ist in der EU, wo natürlich alle Sprachen aller Mitgliedsländer Arbeitssprachen sind und somit natürlich alles in alle Sprachen übersetzt wird. Aber das ist ja bei der NATO nicht.
94 **Sophie Grossalber:** Ja, ich fand auch einfach spannend, dass Englisch und Französisch sind die Arbeitssprachen und dann wurde speziell die Seite dann auch noch auf Russisch und Ukrainisch übersetzt.
95 **Interviewee 2:** Mit denen möchte man natürlich auch kommunizieren. Die NATO ist natürlich interessiert daran, dass die Öffentlichkeit in Russland und in der Ukraine über das Bescheid weiß, ist klar.
96 **Sophie Grossalber:** Genau.
97 **Interviewee 2:** Außerdem ist auch nicht zu erwarten, dass sie die Arbeitssprache Englisch so können wie die Leute in den NATO-Ländern oder … Oder sagen wir mal ja, weil man sich innerhalb der NATO-Länder ja schon auf diese Verkehrssprache oder Arbeitssprache geeinigt hat. Das ist ja mit Russland und der Ukraine nicht so.
98 **Interviewee 2:** Ja, hochinteressant.
99 **Sophie Grossalber:** Mhm, gut. Dankeschön für ihre Zeit.

100 **Interviewee 2:** Gerne.
101 **Interviewee 2:** Alles Gute für die Arbeit.

Experiment Interview 3, October 9th, 2024

1 **Sophie Grossalber:** Gut, erstmal danke fürs Bereiterklären für das Interview. Und nur noch einmal für die Aufnahme, es ist eh okay, dass es dann anonymisiert zitiert wird das Transkript in der Arbeit?
2 **Interviewee 3:** Ja.
3 **Sophie Grossalber:** Okay, dankeschön. Dann fangen wir gleich an. Inwiefern haben sie als Teil ihrer Tätigkeit im Bundesheer mit Übersetzungen zu tun?
4 **Interviewee 3:** Ich bin seit Jänner eingeteilt auf den Arbeitsplatz, Übersetzerin Englisch. Also ja, ich habe auch schon als Praktikantin übersetzt und zwischendurch mit Werkverträgen auch fürs Bundesheer übersetzt. Und ja, jetzt bin ich tatsächlich mal auf dem Arbeitsplatz. verwendet.
5 **Sophie Grossalber:** Okay, haben Sie bereits einmal ChatGPT, DeepL oder ein anderes Tool zur Übersetzung genutzt?
6 **Interviewee 3:** DeepL, ja. ChatGPT hab ich noch nie zum Übersetzen benutzt.
7 Sophie Grossalber: Warum haben Sie sich für eine maschinelle Übersetzung entschieden?
8 **Interviewee 3:** Meistens, weil es schneller geht. Also wirklich: Geschwindigkeit, weil wenn ich dann nur noch überarbeite quasi und halt die ganze Fachtechnologie. Da muss ich ja dann eh noch mal drüber gehen. Aber dass dieses Grundgerüst mal steht, ja die Geschwindigkeit einfach.
9 **Sophie Grossalber:** Und was war Ihr allgemeiner Eindruck von den Übersetzungen?
10 **Interviewee 3:** Die, die bei DeepL rauskommen, meinst du? Es kommt darauf an, also wenn es nur so allgemeinsprachliche Sachen sind, dann ist es sehr gut, aber alles was dann irgendwie so ins Militärische mehr geht, das kann DeepL halt von der Terminologie her einfach nicht.
11 **Sophie Grossalber:** Gut, dann kommen wir zum zweiten Teil, beziehungsweise dem eigentlichen Teil. Ich werde Ihnen gleich drei Übersetzungen vorlegen, eine davon wurde mit ChatGPT angefertigt, eine mit DeepL und eine von einem professionellen Übersetzer mit dem nötigen Fachwissen.
12 **Interviewee 3:** Okay, ja.
13 **Sophie Grossalber:** Ich werde Ihnen allerdings nicht sagen, welche Übersetzung aus welcher Quelle stammt. Die Übersetzungen sind alle derselbe Auszug der NATO-Webseite über das Comprehensive Assistance Package für die Ukraine.
14 **Interviewee 3:** Ja.
15 Sophie Grossalber: Also die Ausgangssprache ist Englisch, alle auf Deutsch übersetzt.
16 **Interviewee 3:** Ok.
17 **Sophie Grossalber:** Und ich würde Sie dann bitten, dass Sie sich Notizen machen und sagen, welcher Text von Ihnen aus DeepL, ChatGPT oder Mensch ist.
18 **Interviewee 3:** OK, ja, passt, gut. Okay.

19 **Interviewee 3:** Hm. Das ist für mich, das ist sicher menschlich. Aber bei den Zweien… Also warte mal. A. B. C. Ich weiß nicht. Ich glaub, A ist DeepL. Und B ist ChatGPT. Aber ich weiß nicht, dadurch, dass ich ChatGPT noch nie verwendet hab, weiß ich nicht wie das übersetzt.

20 **Sophie Grossalber:** Mhm. Okay, gab es bestimmte Gründe für diese Auswahl?

21 **Interviewee 3:** Okay. Also, dass C, das vom Menschen ist, ist weil… Also, ich habe es als erstes eigentlich an dem Punkt da gemerkt, wo Führung, Information, Kommunikation, Computersysteme. Und nicht Kommando, Kontrolle, Kommunikation, weil es ist halt Command and Control und das ist einfach wörtlich übersetzt und das übersetzt man aber eigentlich nicht so wörtlich, wenn man Hintergrundwissen hat. Deswegen. Das Habe ich eigentlich daran… und es ist auch generell, es ist anders formuliert. Also ja, und… Und was ich eh schon gesagt hab mit dem Aufzählungszeichen, also da war ich mir sicher, dass das die Maschine ist. Allein schon, weil da ein Aufzählungszeichen fehlt. Aber ja. Okay. Und bei den anderen Zweien. Also ich hab jetzt voll lange überlegt, weil da einmal ist es als Soldatinnen und Soldaten und einmal ist es als Dienstmänner und -frauen und wahrscheinlich ist es auf Englisch einfach service men. Und das ist halt auch schon wieder so komplett wörtlich übersetzt. Und da war ich mir halt nicht sicher, ob ChatGPT so vielleicht so wörtlich übersetzt, weil ich eher glaub, dass DeepL das schon als Soldaten, aber ich bin mir nicht sicher, keine Ahnung.

22 **Sophie Grossalber:** Mhm.

23 **Interviewee 3:** Und… Genau und dann menschlich war es ja auch Soldaten. Ja, okay.

24 Sophie Grossalber: OK, gut, kommen wir zur Auflösung, Text A ist DeepL, B ChatGPT und C ist Mensch.

25 **Interviewee 3:** Okay, gut.

26 **Sophie Grossalber:** Ja, dass sich die Auflösung mit dem Ranking deckt, wissen wir. Haben Sie sonst noch irgendwelche Anmerkungen zu Ihrem Ranking? Jetzt mit dem Wissen.

27 **Interviewee 3:** Ähm. Also nur grundsätzlich, warum ich mir schon, also bevor ich bis da runter gekommen bin, warum ich mir schon gedacht hab, da im ersten Absatz schon, dass das sicher maschinell ist, ist, dass ich find, dass die Maschinen, also es ist aber eh bei beiden ziemlich ähnlich. Die… Also. Wenn die Sätze so, ich mein, das ist jetzt von Englisch auf Deutsch, aber es ist für mich in beide Richtungen so, wenn es so verschachtelte Sätze gibt, dann... Ja, also ich find die sind immer so umständlich. Die Sätze, die bleiben dann immer in so einer Satzstruktur, die wenn… Wenn das ein Mensch sagen würd, man würd es nicht so komplett verschachteln, sondern man würd versuchen, dass man das irgendwie in eine schönere Form bringt. Und die maschinellen Übersetzungen, finde ich, bleiben immer so. Ja, so, so, so, so. Ein riesengroßer Satz, der über fünf, fünf Zeilen geht. Wenn ich das übersetzen würde, würde ich vielleicht schauen, dass ich das irgendwie schöner formuliere und das ist halt so mit tausend Nebensätzen und … Ich weiß nicht, ich meine, es ist da jetzt bei der menschlichen Übersetzung auch so, aber trotzdem klingt der Satz nicht so umständlich. Oder bei dem weiß ich jetzt nicht, aber so insgesamt. Also weißt du was ich meine? Ich finde es, wenn

ich selber auch mit DeepL zum Beispiel übersetz, dann stört mich das oft, dass wenn wenn ich halt… irgendwas reinkopier, das kommt halt quasi in genau derselben Satzstruktur wieder raus, auch wenn das eigentlich auf der anderen Sprache gar nicht so gut klingt, dann und dann denke ich mir immer, so, okay, wie könnte man das noch so formulieren, dass es nicht so, ja, so anstrengend ist das zu lesen, oder dass man nicht irgendwie, wenn ich das lese, dann muss ich nachher noch mal zurückschauen, so auf was hat sich jetzt der Nebensatz bezogen, oder? Und ja. Das hab ich jetzt sehr schlecht erklärt, glaub ich. Ich hoffe, du weißt, was ich meine. Ja. Und sonst? Na, eigentlich... Ich meine, im letzten Absatz ist halt noch, wo man dann auch wieder stark merkt…

28 **Interviewee 3:** Dass die menschliche Übersetzung halt zum Beispiel eben Klein und Leichtwaffen, die heißen halt so, das ist halt dann ein Fachausdruck, und da ist einfach von Kleinwaffen und leichten Waffen, weil es das halt dann einfach so wörtlich übernimmt, wie es halt drinnen steht. Und ich finde an sowas erkennt man es dann halt auch, weil eben genau DeepL oder auch ChatGPT die Terminologie nicht kann. Und wenn es sich irgendwie so eher beschreibend anhört, dann ist es meistens eben so wörtlich übersetzt und nicht… Nicht das Wort verwendet, was eigentlich in der, in der Zielsprache dann halt der Fachausdruck dafür wär. Genau. Aber... Ja, ich meine so grundsätzlich von der Aussage, also man merkt trotzdem, dass es halt von der, von der Aussage her… Die Maschinen trotzdem ziemlich gut sind, weil verstehen, worum es geht, tut man trotzdem komplett. Also es ist nicht so, das ist jetzt komplett ein Sinn verzehren würde oder so. Jo, ich glaub das waren alle meine Anmerkungen.

29 **Sophie Grossalber:** Ja, das ist mir auch schon aufgefallen, dass DeepL oder ChatGPT einfach die Satzstruktur und die Grammatik übernehmen von der Ursprungssprache. Und das passt dann überhaupt nicht mehr in der Zielsprache.

30 **Interviewee 3:** Voll. Dann kommen da so, so, so, so Sätze raus, wo du denkst: bitte, was? Und ich finde, gerade wenn du als Mensch übersetzt, denkst du ja eher noch, du denkst da noch anders darüber nach, als wenn du das einfach nur liest. Wenn du das jetzt einfach so auf Deutsch liest, erfasst du halt die Information und denkst dir ja, passt. Aber wenn du es übersetzen musst, dann liest du es und wenn es ein langer Satz ist, dann denkst du nochmal so: okay, was sagt das jetzt wirklich aus? Und dann fangt doch eher an zum Umformulieren, weil du dann versuchst okay. Die Aussage muss drinnen bleiben und es soll aber ein schöner, normaler Satz rauskommen und da merkt man, finde ich immer, wie man eigentlich, wenn man es nur in einer Sprache liest, liest man halt über vieles auch drüber, auch wenn es zum Beispiel so blub blub Sätze sind, wo ein Satz voll aufgebauscht ist und sehr wenig Inhalt hat. Und wenn du dann aber das übersetzen sollst und dann mal drüber, also anfängst, darüber nachzudenken, so okay, was steht da eigentlich? Was ist die Aussage, dann merkst du auch, bei vielen Texten merkst du dann, dass das eigentlich, dass da vieles nichts aussagt oder dass man gar nicht weiß, was das aussagen soll, wo man, wenn man es einfach nur in einer Sprache liest, halt wahrscheinlich drüber lesen würde und sich denken würde, aha okay.

31 **Sophie Grossalber:** Gut. Dann ist das hier das Ende des Interviews. Ich sage nochmal danke.

32 **Interviewee 3:** Gerne.

Experiment Interview 4, October 9th, 2024

Sophie Grossalber: So, jetzt erstmal danke fürs Zeit nehmen und nur noch mal für die Aufnahme, es ist eh okay, dass es anonymisiert und dann in der Arbeit zitiert wird?

Interviewee 4: Ja.

Sophie Grossalber: Perfekt. Gut, Frage Nummer 1: Inwiefern haben Sie als Teil Ihrer Tätigkeit innerhalb des Bundesheeres mit Übersetzungen zu tun?

Interviewee 4: Immer wieder. Ich bin an und für sich Lehrerin, aber ich war anfangs als Übersetzerin da auch, hab am Anfang nur übersetzt. Ja, verschiedene Texte nehme an, die Frage kommt noch.

Sophie Grossalber: Ja, haben Sie da jetzt einmal ChatGPT, DeepL oder ein anderes Tool zum Übersetzen benutzt?

Interviewee 4: Ja, hab ich.

Sophie Grossalber: Welches?

Interviewee 4: In erster Linie DeepL. Anfangs einfach nur zur Übersetzung, später dann auch diese Textverbesserungsfunktion.

Sophie Grossalber: Das neue…

Interviewee 4: Das Neue, ja genau. ChatGPT verwende ich nicht zur Übersetzung so gerne eigentlich. Ich hab es zwar auch schon hier und da gemacht, weil ich im Tool drinnen war, aber wenn Übersetzung, nehm ich lieber DeepL.

Sophie Grossalber: Und was waren Ihre Eindrücke von den Übersetzungen?

Interviewee 4: Sie werden immer besser, sie werden immer besser, aber… Ja, es kommt immer auf den Ausgangstext an. Ja, also das ist für mich immer ein Kriterium, wie schaut der Ausgangstext aus, wie schaut das aus, was DeepL ausspuckt und manchmal ist es wirklich gut, manchmal ist es auch viel zu verschachtelt zum Beispiel. Also ich muss es immer nachkorrigieren, aber ich bin halt viel schneller mit DeepL. Aus.

Sophie Grossalber: Das heißt, wenn Sie sich für eine maschinelle Übersetzung entscheiden, würden Sie sagen, es wäre die Zeitersparnis.

Interviewee 4: Das ist mein Hauptgrund.

Sophie Grossalber: Mhm, okay. Kommen wir gleich zum zweiten Teil. Ich werde Ihnen gleich drei Übersetzungen vorlegen, alle derselbe Ausgangstext, der englische Teil über das Comprehensive Assistance Package for Ukraine von der NATO. Und das wurde alles auf Deutsch übersetzt, eins davon mit ChatGPT, einer mit DeepL und einer von einem menschlichen Übersetzer mit dem nötigen Fachwissen. Und was ich jetzt gerne von Ihnen hätte, wäre mir zu sagen, was Sie denken, welcher Text maschinelle oder menschliche Übersetzung ist und warum.

Interviewee 4: Mhm.

Sophie Grossalber: Dann geb ich Ihnen ein bisschen Zeit zum Durchlesen.

Interviewee 4: Okay. Also ich glaube, ich habe jetzt mal nur den ersten Absatz angeschaut. Ich glaube A und B sind die maschinellen und C ist das Nicht-maschinelle. A und B sind sehr ähnlich. Was mir… Wo wirklich nur einzelne Wörter anders sind, zum Beispiel Allianz, Bündnis, Grundlage, Basis, Prinzipien, Grundsätze. Andererseits für das Wort Practices gibt es keine ordentliche Übersetzung. Und da schreiben sie Praktiken, und das hat offensichtlich der menschliche Übersetzer mit Best Practices

übersetzt, was man ja auch eigentlich eher verwenden würde, weil Praktiken ist in diesem Zusammenhang sinnlos.

19 **Sophie Grossalber:** Mhm.

20 **Interviewee 4:** Welches jetzt DeepL ist und welches ChatGPT, kann ich nicht wirklich sagen. Aber es ist sehr, sehr ähnlich. Auch die anderen. Vielleicht ist A eher DeepL, weil‘s eher wörtlicher ist. Ich glaub ChatGPT hat inzwischen die Fähigkeit, kann man das so sagen, auch Synonyme eher sinnvoller einzusetzen. Und auch nicht so am Wort zu kleben wie DeepL das auch tut. Aber ich kann mich auch, kann mich auch irren.

21 **Interviewee 4:** Ich sehe gerade den letzten Absatz an, da steht „abgeschlossenen Treuhandfonds haben die Ukraine“. Das ist in allen drei Texten ein einziger Satz. Huch, ein Absatz, aber das ist… Das wäre so ein menschlicher Fehler. Falls das nicht ein Tippfehler hier im… in dem Text. Ja, es ist auch hier. Der letzte Absatz, „physische Sicherheit und Verwaltung von Munitionslagebeständen“, das würde man viel eher sagen als das, wie es da steht in den maschinellen Übersetzungen. Glaube ich.

22 **Sophie Grossalber:** Okay.

23 **Interviewee 4:** Brauchen Sie alle, soll ich, soll ich auf alle Absätze eingehen oder reicht das derweil einmal?

24 **Sophie Grossalber:** Was Ihnen auffällt.

25 **Interviewee 4:** Ja. Okay, also ich schau mal jetzt gerade den zweiten Absatz an. Allein die Zeitenverwendung. Und die Kollokationen. Hier steht zum Beispiel, „sie leistet Beratung“. Und auch, und da steht „strategische Beratung bereitgestellt hat“, das ist so „provided“, das ist eine klassische Übersetzung. Und im Text C da steht, „Unterstützung bietet“, was eher freier übersetzt ist und auch sinnvoller in dem Zusammenhang und auch „hilft die NATO seit vielen Jahren“. Das haben beide gemacht, aber „über viele Jahre hinweg“ würde man nicht sagen, also das ist eher so eine eigenartige Übersetzung, die eher Richtung maschinell geht. Ja, und auch durch die Syntax, die Wortstellung alleine ist eher die englische, also das klebt wieder eher am englischen Original von demher, was ja im Deutschen jetzt nicht völlig falsch ist. Aber komisch teilweise. Unnatürlich, sagen wir mal. „Durch diese Programme“. Was steht da? „NATO die Kapazitäten gestärkt“, „erheblich gestärkt“. „Erheblich gestärkt“ haben sie alle geschrieben.

26 **Interviewee 4:** Hm. Text C ist sicher nur von einem Menschen gemacht? Oder ist das von einem Menschen nachkorrigiert und maschinell erstellt worden, weil eben auch da trotzdem sehr viele einzelne Phrasen sehr ähnlich sind. In den anderen Texten.

27 **Sophie Grossalber:** Also es wurde eigentlich in Auftrag gegeben, selbst zu übersetzen, aber …

28 **Interviewee 4:** Man weiß ja, wie wir Übersetzer gerne arbeiten. Okay. „Entwicklung von Fähigkeiten“ statt „Fähigkeitsentwicklung“. Okay.

29 **Interviewee 4:** Das ist jetzt wieder ein Text A am besten, der dritte Absatz. Wenn ich mir das anschaue. Der zweite Satz, der dritte Satz, „Die aktiven Projekte des Treuhandfonds konzentrieren sich auf“ und die anderen beiden schreiben „aktive Treuhandfonds-Projekte konzentrieren sich auf“. Das spricht dann wieder eher dafür, dass… Also da ist das wieder besser.

30 **Sophie Grossalber:** Mhm.

31 **Interviewee 4:** Auch wenn es eventuell von der Maschine gemacht wurde. Führung, Information, Kommando, Kontrolle. Führung, Kontrolle, Kommunikation, Computer. Mhm, ich glaub trotzdem dass C menschlich ist, weil einfach auch stark von den… von den Vokabeln her und von den Synonymen abweicht, weil wenn man anschaut den Absatz vier, der nur aus diesem kleinen Satz besteht. Schreibt A „Kommando, Kontrolle, Kommunikation, Computer", B schreibt „Führung, Kontrolle, Kommunikation, Computer" und Text C ist „Führung, Information, Kommunikation, Computersysteme" anders.

32 **Sophie Grossalber:** Mhm.

33 **Interviewee 4:** Reorganisation, Reorganisation, das schreiben… Umstrukturierung ist eher ein Wort, das wir verwenden würden vielleicht im Deutschen als Reorganisation.

34 **Interviewee 4:** Ziviler Mitarbeiter. Ich glaube auch, dass ein Mensch „Mitarbeiter und Mitarbeiterinnen" schreiben würde. Hoffentlich.

35 **Sophie Grossalber:** Mhm.

36 **Interviewee 4:** Ich bin jetzt im Absatz medizinische Rehabilitation. Text C schreibt „ukrainische Soldaten und Soldatinnen", die schreiben hier „entlassene und ukrainische Dienstmänner und -frauen".

37 **Sophie Grossalber:** Mhm.

38 **Interviewee 4:** Was auch komisch ist. Also eher so Richtung Maschine geht eigentlich. So, da steht allerdings im Text A... Text A und C sind eigentlich da sehr… fast deckungsgleich. „Und Soldaten, sowie…" Ja, „Berufsentwicklungsprogramm", das ist komisch. Programm. Da steht „Berufsfortbildungsprogramm". In Text C steht „Programm zur beruflichen Weiterentwicklung", was natürlich, viel natürlicher klingt, sagen wir so.

39 **Sophie Grossalber:** Mhm.

40 **Interviewee 4:** Und „Zivilisten". Schreibt C allerdings auch nur „Zivilisten". Hier schreiben alle drei „Zivilisten". Wo ich, zum Beispiel, „Zivilbevölkerung" schreiben würde. Wenn ich so eine Übersetzung bearbeite. Ok, abgeschlossen.

41 **Interviewee 4:** Ja, den letzten Absatz haben wir eh schon besprochen. Hm, ich versuch noch irgendwie zu überlegen, was ChatGPT oder DeepL ist. Aber ich glaube, ich bleib dabei, was ich glaube, A ist DeepL und B ist ChatGPT. Ja, aber ich bin mir gerade ziemlich sicher, dass C menschlich ist.

42 **Sophie Grossalber:** Zur Auflösung, ja. Text A ist DeepL, B ist ChatGPT und C ist Mensch.

43 **Interviewee 4:** Okay.

44 **Sophie Grossalber:** Haben Sie sonst noch irgendwelche Anmerkungen zu den Übersetzungen?

45 **Interviewee 4:** Außer dass Mensch sehr stark von DeepL beeinflusst ist offensichtlich und wirklich ganz offensichtlich mit dem Programm gearbeitet hat und dass da auch noch Luft nach oben ist, weil gewisse Dinge man trotzdem nicht sagen würde oder schreiben würde. Das ist alles, aber ich habe das Problem in meinem täglichen Umfeld.

46 **Sophie Grossalber:** Ja. Deswegen gibt es das Vier-Augen-Prinzip.

47 **Interviewee 4:** Genau.
48 **Sophie Grossalber:** Gut, dann danke für die Zeit.
49 **Interviewee 4:** Sehr gerne, es war lustig.

Experiment Interview 5, October 9th, 2024

1 **Sophie Grossalber:** Gut. Erstmal danke fürs Zeit nehmen fürs Interview. Und nur noch mal für die Aufnahme, es ist eh okay, dass es dann anonymisiert, transkribiert, in der Arbeit zitiert wird?
2 **Interviewee 5:** Freilich.
3 **Sophie Grossalber:** Dankeschön. Dann fangen wir gleich an. Inwiefern haben Sie als Teil ihrer Tätigkeit beim Bundesheer mit Übersetzungen zu tun?
4 **Interviewee 5:** Manchmal bitten Kollegen und Kolleginnen mich um Hilfe, weil sie ein zweites oder drittes Augenpaar sozusagen brauchen. Sonst nicht bis dahin.
5 **Sophie Grossalber:** Haben Sie bereits einmal ChatGPT, DeepL oder ein anderes Tool wie zum Beispiel Google Translator zur Übersetzung genutzt?
6 **Interviewee 5:** Ja, hab ich.
7 **Sophie Grossalber:** Mhm, welches?
8 **Interviewee 5:** Ganz viel Google translate.
9 **Sophie Grossalber:** Mhm.
10 **Sophie Grossalber:** Für welche Sprachrichtungen?
11 **Interviewee 5:** Vor allem Italienisch, Französisch. Sprachen, die ich noch so halb beherrsche. Und hin und wieder für Russisch oder so Sprachen, die ich gar nicht kenne.
12 **Sophie Grossalber:** Warum haben Sie sich für eine maschinelle Übersetzung entschieden?
13 **Interviewee 5:** Also ich habe Google Translate verwendet, weil ich halt nicht unbedingt…, weil ich nur für mich was übersetzen wollte. Das heißt, wenn ich irgendwelche Nachrichten aus dem Ausland verstehen wollte.
14 **Sophie Grossalber:** Und was war Ihr Eindruck von den Übersetzungen?
15 **Interviewee 5:** Eigentlich habe ich einen ziemlich guten Eindruck bekommen, weil Italienisch und Französisch natürlich auch große, bekannte Sprachen sind und von daher und auch weil ich die Sprache ein bisschen beherrsche und deshalb auch ein bisschen was zum Abgleichen habe. Sozusagen war ich immer bei Google Translate, immer in den letzten Jahren ziemlich zufrieden. Am Anfang nicht, aber jetzt schon.
16 **Sophie Grossalber:** Okay, kommen wir zum zweiten Teil. Ich werde Ihnen gleich drei Übersetzungen vorlegen. Der Ausgangstext ist bei allen derselbe, Ausgangssprache ist Englisch und in dem Text geht es ums Comprehensive Assistance Package für die Ukraine von der NATO. Ich werde Ihnen allerdings nicht sagen, welche Übersetzung ChatGPT, DeepL oder die menschliche Übersetzung ist. Und ich würde Sie dann bitten, dass Sie die Texte dementsprechend in ein Ranking setzen und sagen, welches ist DeepL, welches ist ChatGPT und welches ist menschlich und warum. Aber nehmen Sie sich so viel Zeit, wie sie wollen zum Durchlesen.
17 **Interviewee 5:** Okay. ChatGPT 4 oder 5?
18 **Sophie Grossalber:** 4.

19 **Interviewee 5:** Brauchst du unbedingt eine Antwort?
20 **Sophie Grossalber:** Schon gerne, ja. Okay. Warum haben Sie sich für dieses Ranking entschieden?
21 **Interviewee 5:** Ich bin kein Muttersprachler. Erstens gehe ich davon aus, dass der Übersetzer ein Professional ist, der sich auch auskennt mit militärischen Begriffen. Ich fand den letzten Absatz bisschen detaillierter und spezifischer in den Begrifflichkeiten, deshalb…
22 **Sophie Grossalber:** Mhm.
23 **Interviewee 5:** Habe ich da einen Menschen erwartet und manche Sachen werden, vielleicht, es ist ja ziemlich genau gleich in A und B und werden dann von den Computern nicht unbedingt als Fachbegriff wahrgenommen.
24 **Interviewee 5:** Im ersten Absatz… Hat es bisschen nach einem Giveaway ausgeschaut. Best Practices. Es ist ja nicht abnormal englische Begriffe, englische Idioms in deutschen Texten anzutreffen.
25 **Sophie Grossalber:** Mhm.
26 **Interviewee 5:** Wohingegen der Rechner oder die KI alles zu übersetzen pflegt.
27 **Sophie Grossalber:** Mhm.
28 **Interviewee 5:** Best Practices sind hier „bewährten Praktiken". Ansonsten, es tut mir leid, ich habe sehr wenige, sehr, sehr wenige Hinweise nur für…
29 **Sophie Grossalber:** Mhm.
30 **Interviewee 5:** Für menschliche Kreativität. Ja.
31 **Sophie Grossalber:** Kein Problem. Sonst noch irgendwas, was ins Auge geschlossen ist?
32 **Interviewee 5:** Ja, ich fand es interessant, dass gleich im ersten Absatz. Ähm. Da gibts für dieses Paket drei unterschiedliche Begriffe. „Beistands-", „Hilfs-" und „Unterstützungspaket".
33 **Sophie Grossalber:** Mhm.
34 **Interviewee 5:** Deshalb hatte ich … Da gibt es so ganz kleine, aber immerhin Bedeutungsunterschiede. Und auch „Bündnis" in A und C. Während es in Text B „Allianz" ist und ich bin einfach davon ausgegangen. Aha.
35 **Sophie Grossalber:** Mhm.
36 **Interviewee 5:** Eine KI, egal welche, wird immer auf das Wort „Bündnis" zurückgreifen in einer bestimmten Sprache.
37 **Interviewee 5:** Und das wird dann wohl die Menschliche sein. Und dann habe ich gesehen, dass es eben auch Unterschiede zwischen diesen beiden Texten gibt, aber auch zwischen diesen beiden Texten und tja, ehrlich gesagt, ich sehe qualitätsmäßig, qualitätsmäßig wirklich, wirklich sehr, sehr wenig Unterschiede. Wenn ich ehrlich bin.
38 **Sophie Grossalber:** Mhm.
39 **Interviewee 5:** Kleine Bedeutungsunterschiede vielleicht, aber im Großen und Ganzen ist es dreimal der gleiche Text. So wie ich das einstufen würde.
40 **Sophie Grossalber:** Ja gut, dann kommen wir zur Auflösung. Text A ist DeepL, B ist ChatGPT und C ist der Mensch.
41 **Interviewee 5:** C ist der Mensch, die andere beiden habe ich verdreht.

42 **Sophie Grossalber:** Würden Sie mit dem Wissen jetzt Änderungen im Ranking vornehmen oder haben Sie sonst noch irgendwelche Anmerkungen?

43 **Interviewee 5:** Ich glaub, dass AI sehr wohl und sehr gut für solche Übersetzungen einzusetzen ist. Nur dann, wenn am Ende ein Mensch drüber schaut. Weil der wird tatsächlich sich mit einigen Begrifflichkeiten, und wie die eingesetzt werden, besser auskennen. Vielleicht, dass sich das in Zukunft verändert, vielleicht, dass GPT-5 eine noch bessere Arbeit geleistet hätte in diesem Bereich. Aber AI… fantastisch in Kombination mit einem Paar menschlichen Augen. Das würde ich sagen.

44 **Sophie Grossalber:** Gut, dann sage ich danke für die Zeit.

45 **Interviewee 5:** Bitte gerne.