universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Neural machine translation versus human translation in the field of natural science"

verfasst von / submitted by

Maren Etzlinger, BA

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

Master of Arts (MA)

Wien, 2022 / Vienna 2022

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:	Ass.-Prof. Mag. Dr. Dagmar Gromann, BSc

Table of contents

Acknowledgements

Finishing my degree and writing this thesis while having a job without losing my sanity in the process has been quite a challenge, to put it mildly. There are many people who have played a part in getting me to this point – (extended) family, old friends, co-workers – and although I could not mention every single one of them by name here, I trust that they know that I am grateful to them, too.

First and foremost, I would like to express my gratitude to Ass.-Prof. Mag. Dr. Dagmar Gromann, BSc for her friendly guidance, her helpful suggestions and her professional input throughout the entire process of my writing this thesis.

Next, I want to thank my immediate family; my parents Daniela and Gerald Etzlinger, not only for financially supporting me and enabling me to complete my education, but also for motivating me to keep going, as well as my brother Lukas Etzlinger for sharing my sense of humour and never failing to make me laugh.

Furthermore, I am immensely grateful to my best friend Lisa Jedinger for her emotional support in my many hours of need, for always having a sympathetic ear and for reminding me not to be too hard on myself.

Another "Thank you" goes out to my friend and fellow student Janina Göschl for the ongoing exchange of experiences, thoughts, and ups and downs regarding our respective master's theses.

I am also grateful to my friends Julia Posch and Olivia Su for countless fun (and, at times, not so fun) hours at university and for making my time at the English Department into something special.

Moreover, I want to thank John Williams, Patrick Doyle, Nicholas Hooper, and Alexandre Desplat, whose work has helped me to stay focussed throughout the writing process.

Lastly and most importantly, I am eternally grateful to my boyfriend Thomas Hahnenkamp, not only for showing endless patience while helping me to get my survey and the formatting of this thesis right, but above all for his love, his unwavering encouragement and for always believing in me, especially when I didn't.

List of abbreviations

AEM	Automatic Evaluation Metric
AI	Artificial Intelligence
ALPAC	Automatic Language Processing Advisory Committee
ANN	Artificial Neural Network
BIBD	Balanced Incomplete Block Design
BLEU	Bilingual Evaluation Understudy
BPE	Byte-Pair Encodings
BWS	Best-Worst Scaling
CAT	Computer-Assisted Translation
CBMT	Corpus-Based Machine Translation
CEFR	Common European Framework of Reference for Languages
CNN	Convolutional Neural Network
CUBBITT	Charles University Block-Backtranslation-Improved Transformer Translation
D	DeepL Translator
DDMT	Data-Driven Machine Translation
DMT	Direct Machine Translation
EBMT	Example-Based Machine Translation
FAMT	Fully Automated Machine Translation
G	Google Translate
H	Human
HAMT	Human-Aided Machine Translation
HT	Human Translation
IMT	Interlingual Machine Translation
LISA	Localization Industry Standards Association
M	Microsoft Bing Translator
MAHT	Machine-Aided Human Translation
METEOR	Metric for Evaluation of Translation with Explicit Ordering
MQM	Multidimensional Quality Metrics
MT	Machine Translation
N	No
NLP	Natural Language Processing

NMT	Neural Machine Translation
PC	Professional Context
PNS	Participant of Natural Science
PT	Participant of Translation
RBMT	Rule-Based Machine Translation
RNN	Recurrent Neural Network
RUT	Random Utility Theory
S	SYSTRAN Translate
SAE	Society of Automotive Engineers
SMT	Statistical Machine Translation
TBMT	Transfer-Based Machine Translation
TER	Translation Edit Rate
TM	Translation Memory
TQA	Translation Quality Assessment
WER	Word Error Rate
WMT	Workshop in Machine Translation
Y	Yes

List of figures

List of tables

1. Introduction

The concept of Machine Translation (MT), i.e., translation produced by a computer either without any human intervention or with pre- or post-editing, is being used more and more often in everyday life. The newest type of MT, Neural Machine Translation (NMT), relies on Artificial Intelligence (AI) and deep learning and has been found to achieve very good results that, in some cases, can even be hard to distinguish from a Human Translation (HT). Thus, NMT has become the state of the art, which is reflected both in the tools that are available online for public use and in most research that is conducted within the field of machine translation today (Koehn 2020: 40).

An unpleasant side effect of the increasing quality of MT systems is the rise of the fear that professional translators might no longer be needed in the future. However, the claim that machine translation would be ready to replace human translators within a few years is already several decades old, and so far, it has not come true. Although there are some individual instances in which MT systems have been said to have outperformed human translators (see Popel et al. 2020), it is usually due to specific circumstances (lack of context in isolated sentences, certain language combinations or limited and very basic vocabulary) when these cases do occur. The majority of studies that have been conducted in this respect, however, point towards the assumption that the output of machine translation is still not good enough to make human intervention completely obsolete, especially if specialist translation with subject-specific language is concerned. As Porsiel (2020: 9) states: "Der Mensch wird im Zentrum dieser Prozesskette stehen (müssen), um weiterhin Qualität gewährleisten zu können." ("Humans will (have to) be at the centre of this process chain in order to be able to continue to ensure quality").

With the advent of the internet, the amount of available text has started to grow more rapidly, not only in terms of literary texts or websites, but also as far as highly technical or scientific texts are concerned, and it can be assumed that the need for the translation of said texts has increased as well. Machine translation seems like the perfect way to cope with these rising text volumes, considering that it only takes a fraction of the time a human translator would require. In addition, making use of MT is usually cheaper than hiring a professional translator. But it must be determined whether the results produced by machine translation can achieve the same quality as or maybe even better quality than HT, particularly in the aforementioned cases of specialist texts that deal with complex scientific concepts and use very subject-specific terminology. The aim of this thesis is to shed more light on this matter by comparing NMT to HT, using highly subject-specific texts from the scientific domain of

chemistry and conducting empirical research in the form of Best-Worst Scaling (BWS) as described by Louviere et al. (2015).

The structure of this master's thesis is divided as follows. Chapter 1 presents the research questions and assumptions for this thesis and explains the motivation behind it as well as its objective. The preliminary concepts and theoretical foundations required to lay the groundwork for the empirical part are introduced in Chapter 2. Chapter 3 gives an overview of related work that deals with similar issues and strives to answer similar research questions as this thesis does, and it presents both the research design and the results of those publications. Chapter 4 marks the transition from the theoretical to the practical part, as it outlines the methodology behind the empirical study conducted as part of this thesis. After the presentation of the results in Chapter 5, Chapter 6 discusses them in the context of the research questions and assumptions as well as with regard to the related work. Finally, Chapter 7 provides an overall conclusion to this master's thesis.

1.1. Research questions and assumptions

There are two research questions that this master's thesis will attempt to answer and that thus form the basis for the empirical research to be carried out:

1. How do professional translators as well as natural scientists rate the raw output of neural machine translation in comparison to human translation in domain-specific texts taken from the field of natural science, more specifically, from the field of chemistry, and in the language direction English to German without being explicitly informed about the type of translation?
2. Which of the readily available online neural machine translation tools used in this experiment yields the best results in terms of a numeric rating without any human intervention such as pre- or post-editing?

As can be inferred from the second question, the framework for the two questions will be the same, i.e., the field of natural science (chemistry) and the language direction (English to German) will remain unchanged. Moreover, both questions will be examined under the presupposition that the resulting translations are intended to be used in a professional context, not merely for gisting purposes.

There are basic assumptions that can be made with respect to the questions outlined above. Regarding the first question, it is assumed that the participants of the evaluation will, on average, pick the human translation as the best one more often than the machine translation, i.e., that the human translation will be perceived as exhibiting higher quality than the machine translation. This assumption is based on the fact that although the use of machine translation for texts from the technical or scientific domain has been on the rise, the unedited results of said machine translation process are – in most cases – of insufficient quality for a professional context (Olohan 2016: 49).

With regard to the second question, it is assumed that among the online neural machine translation tools that are going to be used for the empirical part of the thesis (namely DeepL Translator, Google Translate, Microsoft Bing Translator and SYSTRAN Translate), DeepL Translator will generate the most useful translations, i.e., translations that are rated better than those produced by the other MT tools. This assumption is made on the basis of the results of other research that examined the output of different machine translation tools (see e.g. Hidalgo-Ternero 2020, Tavosanis 2019 or Macketanz et al. 2018). It must be mentioned, however, that assumptions such as this one must not be generalised, i.e., the conditions under which the tools were compared must not be disregarded or taken out of context, as aspects such as the type or quality of the source text play an important and decisive role for the quality of the output (Burchardt & Porsiel 2017: 11). Since the papers mentioned before compare neither the exact same constellation of neural machine translation tools nor the same types of texts as this thesis intends to compare, this second assumption is merely a tentative one.

1.2. Motivation and objective

One of the aspects that motivate the writing of this thesis is the fact that natural science is an ever-growing discipline as far as research and publications are concerned. According to a report by the European Commission (2003: 294), "physics, chemistry and basic life sciences" constitute the field which had the highest number of scientific publications in the years from 1995 to 1999 within the then 15 EU member countries. And even today, roughly two decades later, the premise that the natural sciences form the discipline that produces the most research papers and articles still holds true, as can be seen in Fig. 1.

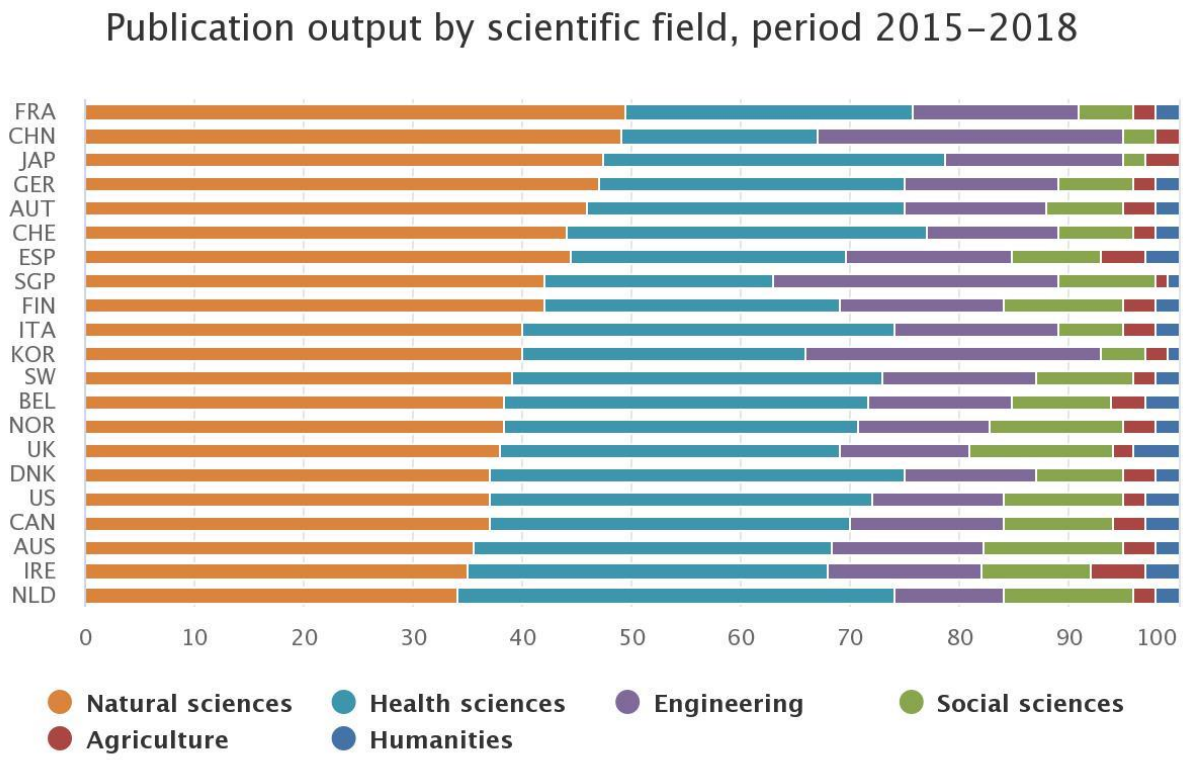


Figure 1: Publication output by scientific field, period 2015-2018 (Rathenau Instituut 2020)

According to the chart in Fig. 1 from the Web of Science database by Clarivate Analytics (Rathenau Instituut 2020), the publication output in the period from 2015 to 2018 was, on average, significantly higher in the field of natural science than in the other fields. This applies to almost all of the 21 countries that have been included in this statistic. In the German-speaking countries Germany (GER), Austria (AUT) and Switzerland (CHE), the percentage of publications in the natural sciences is particularly high, namely 47%, 46% and 44%, respectively. On that basis and in combination with the fact that the volume of text published in scientific fields has been increasing at a yearly rate of ~3% over the past few decades (Bornmann & Mutz 2015: 2217), it can be assumed that the need for interlingual translation within the field of natural science has grown along with the text volumes and will continue to do so in the future.

Another underlying aspect is the use of English as a lingua franca, not only in everyday life, but more specifically in the context of scientific publications. The notion of a lingua franca, i.e., a shared language, for a certain field or discipline has already been made use of multiple times throughout history. For instance, Greek and Latin were the linguae francae for the field of medicine and are still used today among medical professionals around the world to refer to

illnesses, parts of anatomy or other subject-specific concepts. The same has been happening with English for the field of (natural) science (Olohan 2016: 138).

As Montgomery (2009: 13) puts it, "translation will become an even more central communicational aspect" in light of the tendency that scientists around the globe are increasingly relying on the English language for scientific purposes. There is also a growing number of scientific journals that have started to limit their publications to ones written in English (Olohan 2016: 138), which results in an increasing need for translation both into and out of English; into English to allow researchers to actually have their work published in international and renowned scientific journals, and out of English to facilitate the dissemination of knowledge to non-English-speaking countries, researchers, and the general public.

In summary, the objective of this thesis is to make a contribution towards determining whether neural machine translation can be a quick and cost-efficient way to transfer specialist knowledge from English to German. These insights could benefit various groups of people: natural scientists who want or need to access research publications written in English but whose English skills might not be sufficient; computer linguists working on the efficiency of machine translation tools; language service providers or translation agencies planning to incorporate the use of neural machine translation for technical or scientific texts into their business; or students of natural sciences who are reliant on foreign-language specialised literature and who thus require a quick, affordable and at the same time reliable translation of said literature.

2. Preliminaries

This chapter and its subchapters explain the preliminary concepts and theoretical foundations that form the basis for the empirical research conducted as part of this thesis. Chapter 2.1. deals with machine translation, more specifically with its history and development as well as with the different types of machine translation that are available and how they work. Chapter 2.2. is about scientific translation, the particularities involved in the process, and the characteristics of scientific language. Chapter 2.3. gives an overview of different notions of quality in translation, and it lists and describes the most widely known methods of translation quality assessment. Finally, Chapter 2.4. provides more detailed information on the principles and application of best-worst scaling, the research method chosen for this thesis. The concepts introduced in this chapter play an important role because they form the basis for the empirical part of this thesis and because they are necessary in order to understand why the research was designed the way it was.

2.1. Machine translation

Although the meaning of machine translation may seem rather straightforward, the term includes concepts that slightly differ from each other and that will be explained briefly here. In its simplest and most basic definition, machine translation, also sometimes referred to as 'automatic translation', means the use of software for the task of translating natural, i.e., human, languages (Hutchins 2007: 1). This is also known as Fully Automated Machine Translation (FAMT), where humans play no role in the translation process. There is also the concept of Human-Aided Machine Translation (HAMT), where a human may post-edit a machine translation or assist the machine translation process in other ways (Hutchins 2007: 1). The opposite of this would be Machine-Aided Human Translation (MAHT), where the responsibility for the translation resides with a human translator and the machine merely provides assistance in the process, e.g. by means of an incorporated electronic dictionary, terminology database, or editing interface (Kay 1982: 77). This chapter will mainly deal with FAMT rather than HAMT or MAHT.

Even though the focus of this thesis lies on neural machine translation (discussed more closely in Chapter 2.1.3.3.), this chapter will provide an overview of its forerunners, i.e., the various types of machine translation that preceded NMT, as well as their underlying principles

and concepts in order to give the reader a more comprehensive understanding of the developmental process that played a key role in paving the way for the neural approach.

2.1.1. History and development of machine translation

Although it is widely accepted that the first practical steps towards achieving a functioning Machine Translation (MT) system were not taken until the first half of the 20th century, the first ideas that formed the basis for what would later evolve into machine translation date back to the 17th century, namely to the thought of universal and secret languages (Stein 2018: 5). The concept of a universal language was based on the idea that, by means of ontologically unique terms or by developing "a set of the smallest units of meaning" (Stein 2018: 6), it would be possible to construct all thoughts that man could think. A universal language was meant to enable people to communicate regardless of borders or countries and was thus expected to bring peace (Stein 2009: 6). On the other hand, secret languages, as the name already says, were intended to be used for secret and hidden communication by means of codes or encrypted language.

From these rather philosophical ideas of the 17th century, a connection can be established to the 1930s, when two patents for electromechanical multilingual dictionaries were filed; one by the French-Armenian Georges Artsrouni and one by the Russian Petr Trojanskij (Kenny 2019: 429). The latter of the two patents was, retrospectively, more important as it bore closer resemblance to what would later become machine translation. It can be claimed that Trojanskij's device was based on an approach similar to the aforementioned universal language, a set of rules that could be applied to every language and that one merely had to decipher in order to understand the foreign language. Trojanskij intended for the translation process to be performed in three stages (see Hutchins & Somers 1992: 5 and Kenny 2019: 429f.).

In the first stage, the source text would be grammatically analysed by a human operator who only spoke the source language. In other words, by means of a universal set of rules, said operator would provide the individual words of the source text with annotations containing all the words' functions and other grammatical and syntactic information (Hutchins & Somers 1992: 5, Kenny 2019: 429f.). These annotations would also be added to the respective word entry in the dictionary device. In the second stage, the electromechanical device would provide results in the form of target language equivalents of the individual source language words as well as the universal grammatical information that had been added by the first operator (Kenny 2019: 430). In the third and last stage, another human operator, this time a person who

only spoke the target language, would transform the target language words that the machine had output into a proper and correct target text using the grammatical code provided by the first operator (Kenny 2019: 430). Although Trojanskij's proposal can be seen as the basis for machine translation, especially with regard to the three-stage structure, there is one difference in comparison to the machine translation systems which would later emerge: the purpose. While Trojanskij imagined that machine translation be used for specialist texts (Kenny 2019: 430), the reality is that most machine translation systems were (and, for the major part, still are) used for general purposes.

The practical groundwork for machine translation was laid in World War II, when "statistical methods that were processed on computing machines" (Stein 2018: 6) were used to decipher and break German communication that had been encrypted by electromechanical rotor machines. Building on these experiences, Warren Weaver of the Rockefeller Foundation, an American mathematician later known as a pioneer of machine translation, exchanged letters with Andrew Booth, a British engineer and computer scientist, talking about achieving machine translation by utilising computers (Hutchins & Somers 1992: 5). This was the first time that this idea was discussed and is said to have marked the birth of machine translation (Stein 2009: 6). In these letters and in his memorandum from 1949, Weaver not only stated that understanding across countries and cultures was suffering due to the fact that there were so many languages (Kenny 2019: 430), but he also proposed methods and ways to tackle the endeavour of making machine translation possible (Hutchins & Somers 1992: 6).

Consequently, a lot of effort and funding was put into the development of machine translation systems in the 1950s. The focus was not, however, on creating a machine translation system that did not require human intervention, but rather on proving that machine translation was even feasible from a technical point of view (Hutchins & Somers 1992: 6). The first major breakthrough was achieved in 1954 at Georgetown University in Washington, D.C., where the first functioning MT system was introduced, namely in the language combination Russian to English and with a rather small vocabulary limited to 250 words (Hutchins & Somers 1992: 6). Due to the ongoing Cold War, the language pair Russian-English was perceived as crucial during that time. After the demonstration of the first functioning MT system at Georgetown University, the first wave of enthusiasm started in the US, and a popular assumption at the time was that machine translation would be ready for use within five more years (Kenny 2019: 431).

This period of hope and enthusiasm lasted until the second half of the 1960s, when the Automatic Language Processing Advisory Committee (ALPAC) was created and ordered to inspect the progress that had been made so far in the area of machine translation. The infamous

ALPAC report issued in 1966 (ALPAC 1966) smothered all the hope there had been, claiming machine translation to be too slow, too inaccurate and too expensive to be a viable alternative to human translation (Hutchins & Somers 1992: 7). The report also deemed it needless to invest any more time or money into MT research and suggested that researchers instead focus their resources on further advancing machines that assist translators in other ways, like automatic dictionaries (Hutchins & Somers 1992: 7). After this disappointing and discouraging prospect, MT research was almost completely abandoned in the US for more than ten years (Hutchins & Somers 1992: 7).

However, during this period of hardly any research in the US, the topic of machine translation became increasingly interesting in other areas of the world, such as Canada or what was then known as the European Economic Community (Hutchins & Somers 1992: 7). In the case of Canada, the focus was on translation in the language pair French-English, seeing as the country had just become officially bilingual in 1969 (Kenny 2019: 432). The European Economic Community also saw an upturn in translation need as "scientific, technical, administrative and legal documentation [had to be translated] from and into all the Community languages" (Hutchins & Somers 1992: 7). In these areas, the purpose of (machine) translation also began to shift; what had previously been seen primarily as a way of observing the progress other countries or powers were making with regards to science was now becoming a political necessity in terms of civil rights (Kenny 2019: 432). This was also the period in which more effort was put into incorporating linguistics, more specifically semantics, into MT discourse to a greater extent, which, in combination with the increasing use of personal computers, brought new interest and hope to the subject (Stein 2018: 7).

One particularly successful example of an MT system being used in Canada during that time is the TAUM-Météo system, developed by Montreal's research centre and operated by one of the TAUM group's developers, John Chandioux (Poibeau 2017: 87). The system's purpose was to translate weather reports between the languages English and French and not only was it in use for an astoundingly long period of time, namely for approximately 25 years from the late 1970s to the early 2000s, but in its peak times during the 1990s, it was also immensely productive with a translation rate of up to 30 million words per year (Poibeau 2017: 87). There are a number of reasons why the use of the genre of weather reports proved to be so successful in this MT implementation. Firstly, as Kenny (2019: 432) puts it, weather reports are "linguistically simple"; for the most part, they resort to a small range of recurring vocabulary, and they largely stick to the same grammatical patterns and structures. These factors, of course, make it easier for the MT system to yield output that requires barely any post-editing. Another

reason why weather reports lend themselves particularly well to machine translation is that they are of a short-lived nature (Kenny 2019: 432), meaning that they are usually only consumed once and are not intended for being reused or reread. In summary, it can be claimed that the success of the Météo system was crucial for boosting the field of machine translation, considering that it had just suffered a period of hopelessness and abandonment the decade before (Poibeau 2017: 87).

The 1980s brought with them a rise of commercial translation programmes that were being developed for organisations or individual customers particularly in Japan and North America, but the 1980s also saw some translation systems that were intended for personal use on home PCs, which were becoming more and more common during that time (Kenny 2019: 432). However, the output that these translation systems yielded could rarely be used without undergoing substantial post-editing first, which used to be done by hand on paper but was now, in the 1980s, increasingly carried out using computer programmes that were capable of word processing (Kenny 2019: 433). Aside from post-editing, this was also a period in which pre-editing was commonly applied in order to avoid certain mistakes that the translation systems would continue to make. Methods for pre-editing included the implementation of rules, restricting the length of sentences, as well as limiting the vocabulary and using only certain syntactical structures (Kenny 2019: 433). It was in the late 1980s that the proposition was made to draw on statistical methods again and move away from machine translation that was based on syntax and semantics (Stein 2018: 7).

Although machine translation had become available for free and online for the general public by the late 1990s (the first free online MT service was Babel Fish in 1997, see Gaspari & Hutchins 2007: 200), the interest of professional translators in using it kept within limits since most MT systems that were available at the time did not yet yield results that were of sufficient quality to be used for professional purposes and post-editing the output would not have been worthwhile (Kenny 2019: 434). Moreover, the fact that other machine aids that relied on the concept of translation memories (known as Computer-Assisted Translation (CAT) tools) were already abundant on the market (Kenny 2019: 434) is considered another reason why machine translation was paid little attention to. However, the increased use of CAT tools and thus of translation memories in turn facilitated the advancing of automatic translation because large volumes of bilingual, aligned text were now available in a machine-readable format and could thus be used to improve and "feed" Statistical Machine Translation (SMT), which was already on the rise (Kenny 2019: 434).

In the 1990s and 2000s, SMT started to replace Rule-Based Machine Translation (RBMT), which had been prevalent up to that point in time. As mentioned above, the fact that bilingual text was now readily available in large quantities resulted in SMT systems evolving quickly, and thus, SMT was perceived as the standard type of machine translation from the mid-2000s onwards (Kenny 2019: 434). Simultaneously, the demand for translation kept increasing both in the US and in Europe for various reasons: The September 11 terrorist attacks, for instance, resulted in more intense research regarding machine translation from the Arabic language to English in the US; meanwhile the number of official languages of the European Union had once again grown higher as the Union kept expanding (Kenny 2019: 433). In addition, the advancing technological progress necessitated the translation of more and more websites or other online text.

Although the invention of neural networks dates back as far as the 1950s and research on the subject had already been conducted throughout the 1980s (Poibeau 2017: 188), the first successful attempts to incorporate said neural networks into statistical machine translation with data volumes large enough were not made until the early 2010s (Koehn 2020: 39). But even then, the incorporation was a slow process as many researchers had neither access to sufficient computing power nor the necessary experience to use it, as Koehn (2020: 39) states. Around 2013, it was attempted to create MT systems that were entirely neural and that did not rely on statistics whatsoever; this was done by utilising so-called Recurrent Neural Networks (RNNs) and later by using Convolutional Neural Networks (CNNs) (Koehn 2020: 39), artificial networks that are used primarily for image recognition and that are made up of various layers (Koehn 2020: 198), or sequence-to-sequence models, i.e., models "that [take] in a sequence of input bits, then [output] a sequence of output bits" (Neubig 2017: 1). However, the first output that was actually usable was not achieved until several refinements were added. From that point on, it did not take long until neural machine translation was seen as the new standard in terms of machine translation and until researchers almost exclusively focussed on this new type of machine translation (Koehn 2020: 40).

2.1.2. Rule-based machine translation

The earliest type of machine translation, which was also the most common one for several decades, is Rule-Based Machine Translation (RBMT). This approach, adopted in the time after the Second World War and during the Cold War, partially relied on the ideas of encryption and

secret languages that were already mentioned briefly in Chapter 2.1.1., i.e., on the thought that to translate was little more than to decode a message that had previously been encoded in a foreign language (Poibeau 2017: 52) using a set of certain grammatical or syntactical transfer rules – hence the name. However, in order to grasp the full scope of RBMT, it is necessary to go beyond the rather simple-sounding concept of encryption, transmission and decryption. Within the realm of RBMT, there is one categorisation that is illustrated particularly often, namely that of direct translation systems, transfer systems, and interlingual systems (Poibeau 2017: 27f.). All three of these systems form part of RBMT, but they can be distinguished from each other by the level of analytical depth they entail.

The first category, Direct Machine Translation (DMT), comprises systems in which the source language is translated directly, as the name suggests, from the source language to the target language on a verbatim basis without undergoing any transitional phases in between the two (Poibeau 2017: 27). In most cases, this is performed by means of a bilingual dictionary and very basic rearrangement rules that require little to no analysis. The second category, Transfer-Based Machine Translation (TBMT), already involves a deeper level of syntactic or semantic analysis in which the sentence structure of the source text is parsed prior to the translation process. By grouping words into syntactical units and thus refraining from the verbatim approach of DMT, the results are said to be "more idiomatic than with direct translation" (Poibeau 2017: 28). The third and final category, Interlingual Machine Translation (IMT), makes use of a so-called interlingua, a representative intermediate language that is supposed to be a language-independent pivot to which the source language is transferred before being translated into the target language (Poibeau 2017: 28). These three categories and the associated processes have been visualised in the form of Vauquois' triangle, also sometimes known as Vauquois' pyramid, that can be seen in Fig. 2.

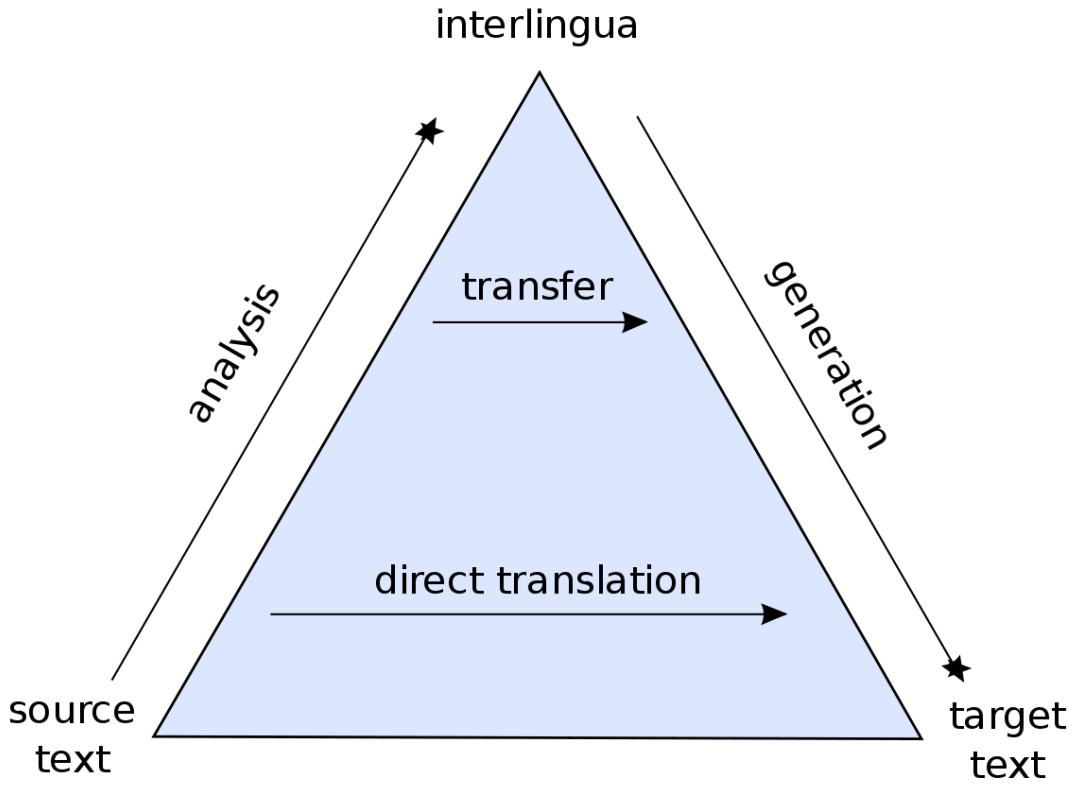


Figure 2: Vauquois' triangle (Wikipedia 2021)

Fig. 2 illustrates the aforementioned different methods or levels of analysis that can be applied within RBMT. As the line entitled "analysis" indicates, the analysis involved in the translation process becomes more extensive and in depth the further one progresses to the top of the triangle. The very bottom of the triangle represents the shallowest level, namely direct translation, which, as mentioned above, requires barely any analysis at all since it uses dictionaries to produce word-for-word translations (Poibeau 2017: 29). Due to the fact that most languages have different characteristics and grammatical rules, the output of these word-for-word translation methods is generally only useful if the two languages are similar and/or if the purpose of the translation is to merely provide the reader with very basic information about the content of the source text (Poibeau 2017: 30).

The middle of the triangle represents transfer-based translation, which forms the basis for most RBMT systems that are in use (Stein 2018: 10). For this method, the sentence is first grouped into fragments that should be treated and translated as one unit (Poibeau 2017: 30). This step already involves morphological and syntactic analysis as well as rules that define how

sentences are constructed in the respective language, for instance that attributive adjectives are placed before the noun in the English language, but after the noun in the Italian language. Other so-called transfer rules deal with semantics and context (Poibeau 2017: 31) and might, for example, prescribe how one word is to be translated in two different ways for two different subject areas. While this transfer-based approach yields better results than the direct one, the disadvantage of these transfer rules is that they only ever apply to two languages, i.e., one language pair (Poibeau 2017: 31). In turn, this means that each language pair requires its own set of rules, making it a tedious and time-consuming endeavour.

This problem is precisely the one that the very top level of the triangle, the interlingua, intends to solve. The objective of the interlingua method is to transfer the content from the source language to a representative and independent pivot language referred to as interlingua and to subsequently transfer the content from the interlingua into multiple different target languages at once without the need for language-specific adaptations (Poibeau 2017: 32). This approach sounds good in theory, but seeing as it would be necessary for the machine to fully comprehend abstract meanings and concepts as well as "to produce linguistically valid sentences in the different target languages" (Poibeau 2017: 32), the reality is that this matter still remains too complex and challenging for current systems and technologies. Moreover, Stein (2018: 10) considers the idea of creating an interlingua that can be used to represent all languages in their entirety of rules and specificities to be utopian. In summary, it can be said that RBMT systems were and continue to be refined, intricate systems that still enjoy great popularity, but due to the large number of complex rules they involve, maintaining them is often not worthwhile (Poibeau 2017: 50).

2.1.3. Corpus-based machine translation

Due to its high level of maintenance and in light of the advent of the internet, in the mid-1980s, RBMT thus gradually came to be replaced by methods belonging to the category of Corpus-Based Machine Translation (CBMT), also sometimes referred to as Data-Driven Machine Translation (DDMT) (Kenny 2019: 435). CBMT comprises Example-Based Machine Translation (EBMT), Statistical Machine Translation (SMT), which became the standard approach in the late 1990s and early 2000s, as was already briefly mentioned in Chapter 2.1.1., as well as Neural Machine Translation (NMT) (Kenny 2019: 435). As both the terms CBMT

and DDMT suggest, this type of machine translation relies on the utilisation of (textual) data, which is usually stored in corpora or, in most cases, parallel corpora. These parallel corpora are collections of texts and their translations, i.e., bilingual texts that are paired on a sentence-level based on their equivalence in meaning. Each of these bilingual sets is called a pair, and the process of pairing the sentences that belong together is known as sentence alignment. As Poibeau (2017: 91f.) explains, "[an] aligned parallel pair of texts is called a [bitext], from bilingual text", and needless to say, the availability of these so-called bitexts has been steadily increasing ever since the internet was invented.

In the form of so-called Translation Memories (TMs), i.e., digital storage units that contain previously translated segments and can be accessed and browsed for purposes of research and uniformity, parallel corpora have become an indispensable resource and tool for human translators (Poibeau 2017: 92), but they are also extensive data pools and thus an excellent knowledge reservoir which can be used to feed and train MT systems. The MT systems, in turn, can analyse said data and use the resulting findings in different ways. On the one hand, a system can examine and evaluate the existing data, i.e., the translations, and derive general assumptions or rules from it that can be applied again in the future to produce new translations. This process is known as Example-Based Machine Translation (EBMT), where the existing data is perceived as examples for future use (Poibeau 2017: 93). On the other hand, systems are also able to make predictions in the form of statistical models on the basis of the input data, which is known as statistical machine translation. As has already been mentioned, the number of bitexts has been growing rapidly, which is beneficial for the creation of MT systems; the more data there is for the systems to process and analyse, the better and the more reliable their performance and their output becomes (Poibeau 2017: 95).

2.1.3.1. Example-based machine translation

The reason for the advent of EBMT was the fact that RBMT systems, the type of MT that was most common at the time, tended to become increasingly complex regarding their maintenance. On top of that, the RBMT systems failed to yield a result if only an ever so small part of the source language sentence could not be analysed successfully (Poibeau 2017: 109). As an analogy to the procedure mainly used by human professional translators to translate a sentence in fragments as opposed to analysing the entire sentence before starting to translate it, Japanese computer scientist Makoto Nagao suggested that MT systems should use existing translation

units from parallel corpora instead of adding so many new transfer rules to RBMT systems (Somers 1999: 116). An EBMT system usually works in the following three steps (Poibeau 2017: 110). Firstly, it browses the aligned pairs of bitext the given parallel corpus contains and attempts to identify pieces of the source language sentence it was ordered to translate. Secondly, the system looks for the respective target language equivalent that is stored in the aligned pair. Finally, it attempts to merge the single fragments of the target language it has identified into a new grammatically, syntactically and semantically correct sentence in the target language. Fig. 3 visualises the process in a very simplified manner, using an imaginary situation in which the sentence 'Training is not the solution to every problem.' is to be translated from English into French by an EBMT system.

Ex1	*Training is not the solution to everything.* *La formation n'est pas la solution universelle.*
Ex2	*Training is not the solution to all parenting struggles* *La formation n'est pas la solution à toutes les difficultés rencontrées par les parents.*
Ex3	*There is a solution to every problem.* *Il y a une solution à tous les problèmes.*
Ex4	*There is a spiritual solution to every problem.* *Il y a une solution spirituelle à tous les problèmes*

Figure 3: Automatically extracted sentences from an imaginary bilingual corpus in order to translate the sentence 'Training is not the solution to every problem' (Poibeau 2017: 112)

As can be seen in Fig. 3, different units of the source language sentence to be translated can be found in different example sentences from the bilingual corpus, e.g. 'Training is not the solution' occurs in Ex1 and Ex2, while the fragment 'to every problem' can be found in Ex3 and Ex4. The EBMT system compares all of these occurrences and is able to determine that the target language phrase 'La formation n'est pas la solution' is the one that Ex1 and Ex2 have in common, which is why it must be the French equivalent to the English 'Training is not the

solution'. Consequently, the system deduces that 'à tous les problèmes', which occurs in both Ex3 and Ex4, must be the equivalent for the English unit 'to every problem'. As pointed out above, this example makes the process look quite uncomplicated, but the reality is not as simple. Among other things, the problems in practice are that the sentences to be translated are usually longer and more complex than the one used in this example and in most cases, the sequence of matching words that is found in the corpus is only very short (Poibeau 2017: 113). In addition, it can be difficult for the system to select the adequate translation if multiple different target language equivalents are available for the same source language unit. The process of merging the various target language units into a complete sentence can also be a challenge "since fragments often overlap or are not fully compatible with each other" (Poibeau 2017: 114).

Due to these issues, it is necessary to step away from the approach of merely searching for word strings and instead turn to what is known as translation by analogy. In this process, the goal is not to look for the very same text fragment that needs to be translated but rather for phrases that are to some extent analogous to the respective source phrases (Poibeau 2017: 114). These analogies can be determined on a rather superficial level by comparing strings of characters or words, but the techniques that yield the most promising results take place on a deeper level. One such technique utilises linguistic tagging, also known as part-of-speech tagging (Poibeau 2017: 115), i.e., marking phrases as belonging to a specific word class, e.g. tagging the word 'solution' as being a noun or 'is' as being a verb. This pre-processing procedure provides bitexts with extra information that the machine can use to compare the bitext and search for target language equivalents more efficiently. In essence, this approach bears resemblance to the way Trojanskij imagined his invention of an electromechanical multilingual dictionary to be used in the 1930s (explained in Chapter 2.1.1.). Since this linguistic tagging approach also includes transfer rules that are responsible for considering the additional information during the creation of the target sentence, EBMT was often perceived to be "a bitter rival to the existing paradigm [of rule-based approaches]" (Somers 1999: 137) in its beginning stages. Another technique for achieving translation by analogy is to make the EBMT system compare linguistic, e.g. syntactic, structures, which would theoretically enable the system to find examples that have the same meaning but different syntactic structures (Poibeau 2017: 116). In practice, however, this approach barely results in any advancement, considering the fact that it is hard to find sufficient phrase-based examples (Poibeau 2017: 116).

It could be determined that the use of EBMT was most successful in smaller sublanguages (Somers 1999: 118) that share a limited range of vocabulary and that make use of regular jargon and wording patterns. One such example would be the sublanguage of computer documentation, where EBMT was able to achieve sufficient coverage due to the fact that the sentences within the sublanguage largely adhere to the same phraseology and terminology (Poibeau 2017: 119). However, attempts to transfer said coverage to non-sublanguage texts remained futile even for highly regular texts. Thus, while it was deemed to be an interesting approach, EBMT was expected to serve the purpose of a mere modular component of other systems that relied on more complex concepts (Poibeau 2017: 119). In the late 1990s, it still seemed to be certain that EBMT would remain the focus of MT research, not necessarily as a rival to RBMT, but rather as an enhancement and sometimes an alternative to other MT systems (Somers 1999: 150). However, only a few years later, namely by the mid-2000s, EBMT had already been upstaged by its CBMT relative statistical machine translation.

2.1.3.2. Statistical machine translation

Although SMT research had already started in the late 1980s by an IBM Research team and continued in the 1990s (Koehn 2010: 17), it was not until the 2000s that it became the new state of the art in the field of machine translation. Like EBMT systems, SMT systems also draw on data from parallel corpora. The difference, in a nutshell, is that instead of creating a target language sentence using previous examples from corpora, an SMT system makes probabilistic calculations based on the data from corpora in order to predict which of the target language sentences it can create is most likely to be the correct one (Kenny 2019: 435). These calculations can be performed on the basis of models that belong to different categories. The two most widely known categories are word-based models and phrase-based models; the paragraphs below provide a general, simple and superficial description of these two types of models. For a more in-depth and, above all, a more mathematical explanation of SMT, its models, and its functionalities, the reader is recommended to refer to Koehn (2010).

Before word-based and phrase-based models are explained, however, it is necessary to first talk about word alignment and the challenges it entails. Earlier in this chapter, the term ‘sentence alignment’ was already mentioned; it is the process of pairing a source language sentence with its equivalent sentence in the target language. Word alignment is the same process, only at the level of individual words. Although it sounds simple in theory, there are

some pitfalls involved in this endeavour. While it is fairly common that one sentence in the source language corresponds to one sentence in the target language, this is usually not the case with words. Especially when agglutinative languages, such as Turkish or Japanese, or languages that heavily rely on compounding, such as German, are involved, it is likely that it is not a one-to-one alignment, but rather that "one word from the source or target language corresponds to 0, 1, or *n* words in the other language" (Poibeau 2017: 122). In this context, *n* represents the number of consecutive words that are grouped in one unit which corresponds to one word in the other language. These units of word strings are also known as '*n*-grams' and in the SMT field, these *n*-grams are sometimes referred to as 'phrases' (Kenny 2019: 436). Fig. 4 serves as an illustration of what word alignment looks like in a sentence where each word is represented by exactly one word in the other language.

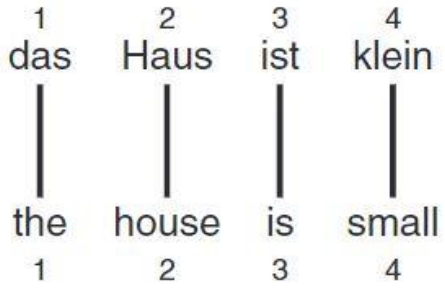


Figure 4: Word alignment in a sentence where each word in one language is represented by one word in the other (Koehn 2010: 84)

In the rather simple case depicted in Fig. 4, it can be clearly recognised that there is one target language equivalent for each of the source language words. While this example illustrates word alignment in a very straightforward manner, it is needless to say that it hardly constitutes the norm, seeing as the example sentence is extremely short and simple, features no modifiers, compounds or relative clauses and is a declarative sentence, i.e., it is not a negation or question. Fig. 5 depicts another rather simple sentence, but this time, not every word in one language has one equivalent in the other.

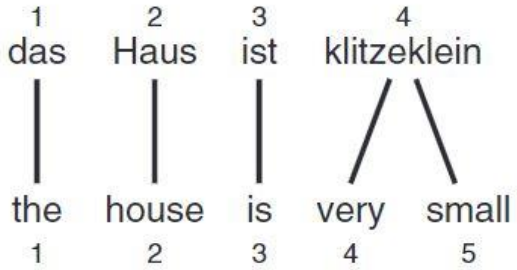


Figure 5: Word alignment in a sentence where a word in one language is represented by several words in the other (Koehn 2010: 85)

The German sentence in Fig. 5 features the word ‘klitzeklein’, which was translated as and aligned with ‘very small’ in this example. Therefore, two words were used in English to convey what was expressed in one word in German. Fig. 6 shows word alignment in a more complex sentence pair.

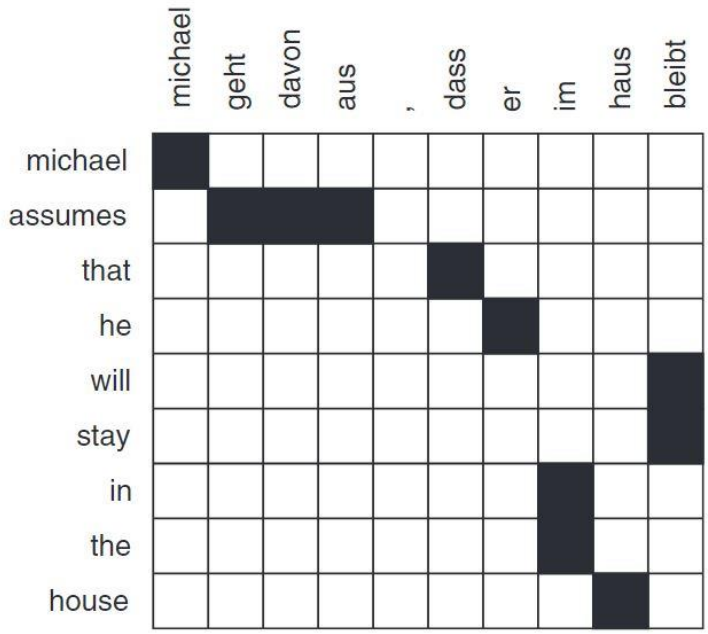


Figure 6: Word alignment in a subordinate clause (Koehn 2010: 114)

Fig. 6 shows a so-called trigram in a translation, i.e., one English word (‘assumes’) that is aligned to a three-word unit, or trigram, in German (‘geht davon aus’). Moreover, there are two so-called bigrams, a two-word unit (German ‘im’ – English ‘in the’; German ‘bleibt’ – English ‘will stay’), and there is also one element that is not aligned to anything in the other language, namely the comma in the German sentence (Koehn 2010: 114). Although the process

of word alignment is already more complex in this sentence pair, it is still possible without any obstacles. Fig. 7, on the other hand, gives two examples of what a problematic case of word alignment might look like.

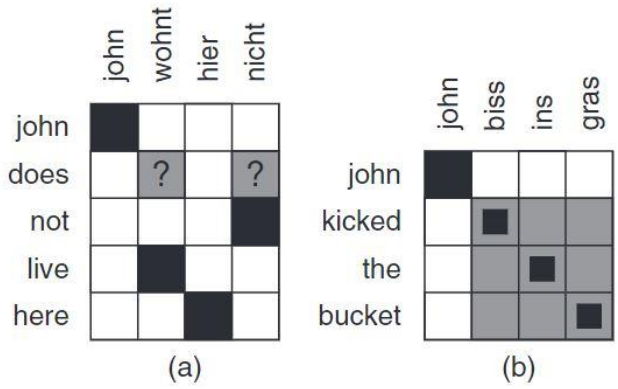


Figure 7: Problematic word alignment (Koehn 2010: 114)

Fig. 7 shows two sentence pairs in which the process of word alignment raises issues. First, there is the sentence pair (a), which is a negation. Aside from a few exceptions, negating an English sentence requires the function word ‘do’ or ‘does’, which carries no meaning in German. Therefore, the question arises whether ‘does’ should be aligned with ‘wohnt’ (‘lives’) “since it contains the number and tense information for that verb” (Koehn 2010: 114) or whether it should belong to ‘nicht’ (‘not’) since the negation is the only reason why ‘does’ is even part of the sentence in the first place. The third solution would be to leave it unaligned. All three of these solutions are valid and can be argued for (Koehn 2010: 114). The other problematic sentence pair, (b), features an idiom. Since both the German idiom ‘ins Gras beißen’ and the English idiom ‘to kick the bucket’ mean the same thing, namely to die, it might seem tempting to just align ‘kicked’ with ‘biss’ (literally ‘bit’), ‘the’ with ‘ins’ (literally ‘in the’) and ‘bucket’ with ‘Gras’ (literally ‘grass’). However, apart from this particular idiom, ‘bucket’ will never be an appropriate equivalent for the German word ‘Gras’. Therefore, this situation requires what is known as a “phrasal alignment” (Koehn 2010: 115), which means that these three-word strings in the respective languages will be treated as one phrase that cannot be broken down into smaller bits.

Now that the principle of word alignment has been outlined in simple terms, it is time to look at the way word-based SMT models work. Although word-based models have been

outdated for quite some time, some of their basic approaches and methods are still applied in more modern systems. The underlying principle for word-based models is mere lexical translation, i.e., the translation of individual words as they would appear in a conventional bilingual dictionary (Koehn 2010: 81). For a very simplified example, Koehn uses the German word ‘Haus’ (2010: 81ff.). There are many possible translations for this word in the English language, for instance ‘house’, ‘building’ or ‘home’, among others. This is the point at which statistics enter the process. In a hypothetical parallel corpus containing the source language German and the target language English, one could count how often the German word ‘Haus’ is translated as ‘house’, how often as ‘building’, and so on. Fig. 8 depicts the hypothetical count which Koehn uses for his exemplary explanation.

Translation of *Haus*	Count
house	8000
building	1600
home	200
household	150
shell	50

Figure 8: Hypothetical counts for different translations of the German word 'Haus' into English (Koehn 2010: 82)

As can be deduced from Fig. 8, in total there are 10,000 occurrences of the German word ‘Haus’ in Koehn’s hypothetical parallel corpus. 8,000 times, it was translated as ‘house’, 1,600 times as ‘building’, 200 times as ‘home’, 150 times as ‘household’, and 50 times as ‘shell’. Using these counts, it is possible to calculate what Koehn calls a “lexical translation probability distribution” (2010: 82), i.e., a value between 0 and 1 which indicates how likely it is that the respective target language candidate is the correct one. The closer the value is to 1, the more common the translation. The closer it is to 0, the less likely the translation is to be correct. The total of all values for one source language word must always equal 1. The most intuitive and uncomplicated way of determining said value for each target language candidate is to divide the count number of the candidate by the total number of source word occurrences (Koehn 2010: 83). In Koehn’s example, this ratio calculation would result in a value of 0.8 (8,000/10,000) for the translation ‘house’, 0.16 (1,600/10,000) for ‘building’, 0.02

(200/10,000) for ‘home’, 0.015 (150/10,000) for ‘household’ and 0.005 (50/10,000) for ‘shell’. As is evident both from the absolute counts in Fig. 8 and from the probability distribution calculated from these counts, ‘house’ is the most common translation for the German word ‘Haus’ in this example and the SMT system will thus most likely use this option. Once again, however, it must be stressed that this example is an extremely simplified illustration of the method, and it does not take context or other knowledge about the texts into account (Koehn 2010: 82).

All of the initial models that were developed by the aforementioned IBM Research team, which are five models in total, are word-based models, but they were already created under the premise that there was not only one-to-one correspondence, but that there could also be one-to-zero or one-to-*n* correspondences. The five models all suggest different improvements, aiming to enhance the system’s ability to cope with expressions that contain more than one word (Poibeau 2017: 132). However, the details of these IBM Models will not be discussed further in this thesis.

Word-based models are more prone to failing when there is no one-to-one correspondence in terms of word alignment, i.e., when the word count between source language and target language differs and when translation must thus take place beyond the word-for-word level (Poibeau 2017: 147). This is often the case when sentences become longer, and the syntactical structure becomes more complex. Phrase-based models can be used where these limits of word-based models are reached. As the name suggests, the phrase-based approach does not operate at the word level, but rather on the basis of phrases, also sometimes referred to as segments or *n*-grams. By translating strings of words instead of individual words, ambiguous formulations can be cleared up. Moreover, it is possible for the system to “learn longer and longer useful phrases” (Koehn 2010: 128), provided that the used corpora are large enough. Phrase-based models not only allow for one-to-*n* alignments, but also for *m*-to-*n* alignments, meaning that “any number of words from the source language corresponds to any number of words in the target language” (Poibeau 2017: 148). Fig. 9 depicts such a case of *m*-to-*n* alignment.

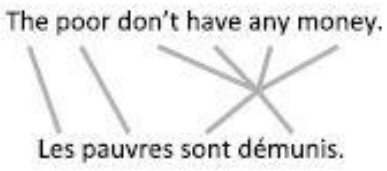


Figure 9: Example of an *m*-to-*n* alignment that would be impossible to obtain from a word-based model (Poibeau 2017: 149)

In the case illustrated in Fig. 9, the source language phrase has $m = 4$ words (if 'don't' is counted as one word) and the aligned target language phrase has $n = 2$ words. This alignment would not have been possible with the original IBM Models, since these models always had one single source language word as the alignment basis. As already briefly mentioned in a previous paragraph, phrases are strings of words that are to be viewed as one unit. The process of segmenting sentences into phrases does not necessarily adhere to any linguistic patterns or rules; theoretically, any string of contiguous words could be seen as a unit (Koehn 2010: 127). Fig. 10, for instance, shows an example where the segmentation has a motivation that is not a linguistic one.

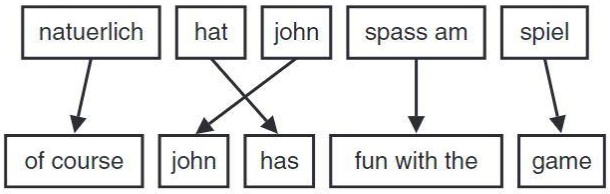


Figure 10: Unusual segmentation in a phrase-based alignment (Koehn 2010: 128)

In the sentence depicted in Fig. 10, the segmentation appears quite random at first glance. From a linguistic point of view, it might seem more natural to group the prepositional phrase 'am Spiel' ('with the game') together into one segment instead of 'Spaß am' ('fun with the'). However, since prepositions are often difficult to translate without context, and since translating the German preposition 'am' with 'with the' is fairly unusual, but in this case correct, it is much more useful for the SMT system to learn that it should translate the phrase 'Spaß am' as 'fun with the' (Koehn 2010: 128). If the phrase were 'am Spiel', the system might, for instance, translate it as 'at the game' or 'on the game', which would not be a good equivalent in this case. Based on these alignments, different phrase pairs can be extracted. Table 1 lists phrase pairs that are possible for the example depicted in Fig. 10.

Table 1: Extracted phrase pairs

German	English
Natürlich	Of course
Natürlich hat John	Of course John has
hat	has
John	John
hat John	John has
Spaß am	fun with the
Spaß am Spiel	fun with the game
Spiel	game
hat John Spaß am Spiel	John has fun with the game
Natürlich hat John Spaß am Spiel	Of course John has fun with the game

The potential phrase pairs in Table 1 range from single words to strings of several words all the way up to the entire sentence. The shorter phrases will prove to be useful since they appear more often in the corpus and will thus be helpful for the translation of entirely new sentences, while the longer phrases are valuable because they provide a lot more information on the context and also facilitate the translation of longer text segments at once (Koehn 2010: 131). After the translation process, phrase-based SMT also includes a reordering process in which phrases are moved around in order to form a grammatically and syntactically correct target language sentence. For this step, it is necessary to calculate how many words are between the original position and the final, desired position of a phrase. The further apart these two positions are, i.e., the more words are skipped between them, the more "expensive" and thus the less likely the reordering becomes (Koehn 2010: 129f.). The reordering process is therefore not something the system learns from the available data, but something that is imposed by means of a predefined model.

In order to produce a translation, any SMT system, be it word-based or phrase-based, needs two more things: a word alignment model and a language model. The concept of word alignment has already been discussed in an earlier paragraph and forms the basis for alignment models. These alignment models contain an alignment function, i.e., a mapping between the source and target language sentences which provides the system with information on the original and target positions of each word within the sentence (Koehn 2010: 84). Of course, these word alignments are initially regarded as hidden variables. In other words, the system must find these alignment functions somehow. This is achieved by means of a specific learning method, the so-called "expectation [maximisation] algorithm" (Koehn 2010: 88f.), in which

equal probabilities are assigned to each possible alignment in the first instance. Then, the probabilities must be tuned, i.e., individual possibilities are weighted with higher or lower values according to the preferable option. This means that alignments which correspond to the aforementioned assumption that 'house' is the most likely translation for 'Haus' are favoured over other alignments and therefore receive a higher probabilistic value.

The other prerequisite, the language model, is responsible for guaranteeing that the output of the SMT system is fluent and correct with regard to word order and word choice (Koehn 2010: 181). Essentially, it calculates how likely a native speaker of the respective language would be to speak the exact target language sentence that the SMT system has created. In contrast to other models in the SMT process, the language model relies exclusively on monolingual data as opposed to bi- or multilingual data. The model is trained to prefer syntactically correct sentences over syntactically incorrect ones, even if the two might feature the same words. As an example for such cases, Koehn (2010: 181) uses the two phrases 'the house is small' and 'small the is house'. Needless to say, a well-trained language model should assign a higher probabilistic value to the first option. But a language model is not only useful for word order. It can also be helpful for the right choice of a word if multiple translations are available for one source language word. Of course, the aforementioned lexical translation probability already gives an estimation of what the most likely translation will be, but in certain contexts, a different translation might be the correct one (Koehn 2010: 181f.). For instance, although the lexical translation probability has determined that 'house' is the most common translation for the German word 'Haus', the language model can overrule this aspect by telling the system that 'I am going home' would be a much more authentic translation than 'I am going house' (Koehn 2010: 182). Since there are so many components involved in the entire SMT process, SMT systems can use the concept of so-called log-linear models, which make it possible to give certain models or elements within the system more influence or weight than others (Koehn 2010: 137).

Although the two SMT models that have been outlined in this chapter are the most widely known ones, they are certainly not the only ones. There are a multitude of different ways in which SMT can be trained and tuned, and a number of aspects that can be used to enhance and refine the results. However, the main principles behind the functionality of SMT have been discussed in a simplified manner. Finally, this chapter shall take a brief look at the challenges and limitations that remain present in SMT to this day.

One of the main challenges in SMT is that of language diversity. It is evident that SMT and the steps it requires, i.e., sentence or word alignment and the identification of target language equivalents, are facilitated if the two languages involved have a linguistically similar structure (Poibeau 2017: 164). For instance, SMT between two Romance languages will usually be easier and more successful than SMT between a Germanic language and a Sino-Tibetan language. But the issue of language diversity varies not only between different language pairs, but also between the language direction within the same language pair. For example, SMT from German into English produces significantly better results than SMT from English into German (Poibeau 2017: 164). Firstly, this is due to the fact that German has a much wider range of morphological variation than English. Secondly, German uses many (partially complex) compound words, and while the SMT system can recognise and decompose said compounds during translation from German to English, it is much more difficult for the system to generate compounds on its own if the language direction is switched (Poibeau 2017: 164). In cases where the two languages have very different linguistic structures, the best results will likely be achieved by means of a hybrid of SMT and a linguistic component that can consider special characteristics of the respective language (Poibeau 2017: 165).

Another issue in SMT is that of rare languages and the subsequent need for a pivot language. As the name of SMT's superordinate category, namely corpus-based/data-driven MT, suggests, SMT systems rely on a vast amount of data, ideally in the form of parallel corpora. These amounts of data are largely collected online, i.e., from texts that are available on the internet. The internet, however, is dominated by a rather small number of languages, as Fig. 11 illustrates.

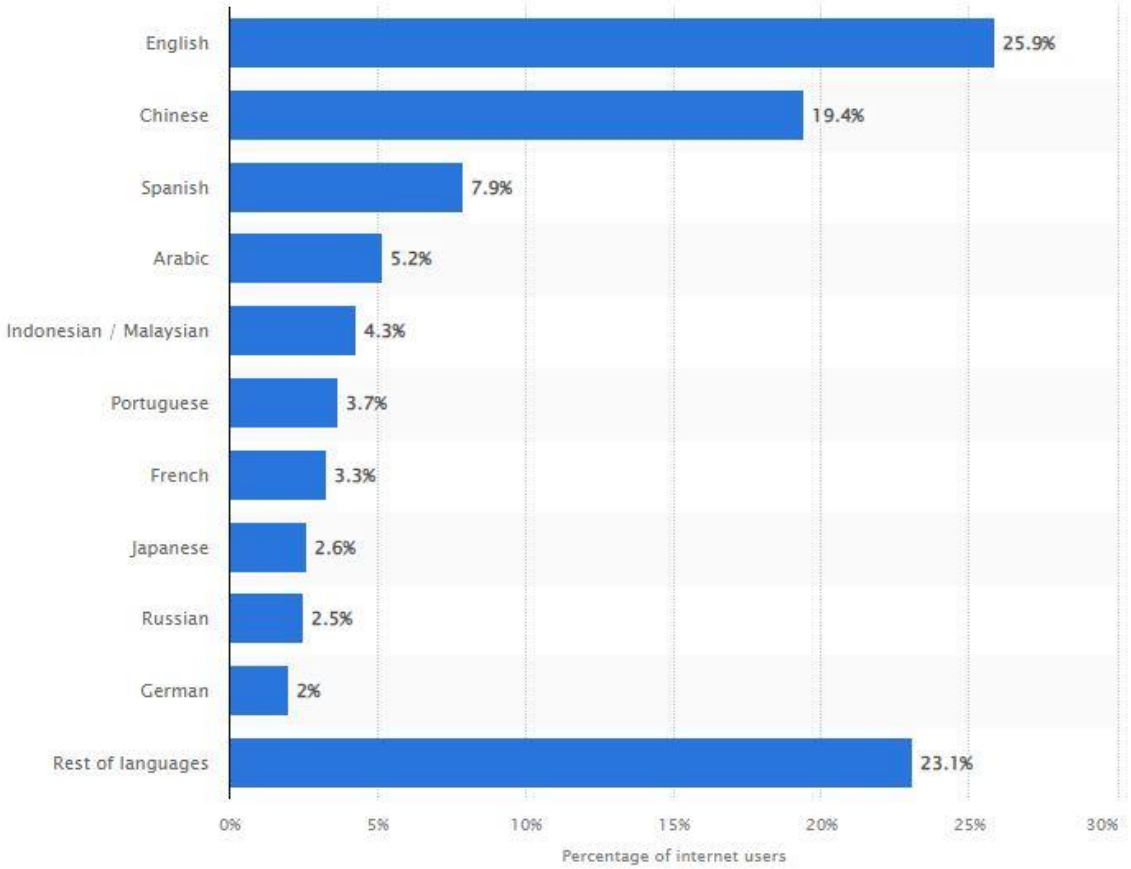


Figure 11: Most common languages used on the internet as of January 2020, by share of internet users (Statista 2020)

Fig. 11 shows the languages that are represented online most often based on the percentage of users. It can be concluded that over half of all internet users (53.2%, to be precise) are represented by only three languages: English (25.9%), Chinese (19.4%), and Spanish (7.9%). For many of the less dominant languages, the data volumes that are available online are simply not enough to achieve good SMT output (Poibeau 2017: 166). Between this circumstance and the fact that the translation quality suffers tremendously "if one of the languages (source or target) is not English" (Poibeau 2017: 166), it goes without saying that an SMT translation between two non-dominant languages, e.g. German and Greek, will thus be considerably more difficult and yield a less satisfying and lower-quality result than an SMT translation involving two, or at least one, dominant language/-s, e.g. English and Spanish. Two different strategies can be applied in order to tackle this problem.

One is to feed the SMT system monolingual data instead of bilingual data, but the results that are achieved with this technique are still considered unsatisfactory (Poibeau 2017: 166). The more common strategy is to make use of the concept of a pivot language, i.e., an interim language to which the source text content is translated first before it is translated from said pivot

language to the actual target language. However, this method can also lead to ambiguities and mistranslations. One example for such translation errors was mentioned by the French researcher Frédéric Kaplan in one of his blog entries (Wordpress 2014): Using Google Translate, which was still using the statistical approach at the time, he translated the French sentence 'Cette fille est jolie' ('This girl is pretty') into Italian, and the output he received was 'Questa ragazza è abbastanza' (literally 'This girl is quite'). This error occurred due to the fact that Google Translate used English as a pivot language and wrongly considered 'pretty' to be an adverb rather than an adjective. Another flaw of this pivot language method is that it often fails to recognise idioms and translates them word by word. Moreover, the use of English as a pivot language leads to English becoming even more dominant, and consequently, the research effort that should be invested in language diversity is curbed yet again (Poibeau 2017: 168).

In summary, it is safe to claim that statistical machine translation has seen undeniable progress and success since its advent in the late 20th century, not least because computational power has been increasing, but also due to the growing amount of bitexts that are available on the internet. Although SMT continues to deal with various struggles to this day, it still shows steady performance, and in combination with other systems or devices, called a hybrid approach, its shortcomings might be improved.

2.1.3.3. Neural machine translation

Neural Machine Translation (NMT), which has been the new state of the art in the field of machine translation for the past approximately six years, is a corpus-based machine translation approach which relies on deep learning and the use of Artificial Neural Networks (ANNs). As briefly mentioned in Chapter 2.1.1., the invention of neural networks dates back to the 1950s and research efforts in this regard were already made in the 1980s (Poibeau 2017: 188). However, the computational power that was available at that time was insufficient for the complexity of the task. Even today, it is still not unusual for the training phase of NMT systems to last for several days, even with the implementation of certain techniques that are intended to significantly speed up and facilitate the procedure (Poibeau 2017: 188). While the neural approach is related to SMT in that it also belongs to the category of corpus-based or data-driven machine translation, i.e., it makes use of large parallel corpora, it differs from SMT in that it consists of "a single big neural network […] that is designed to model the entire MT process" (Luong 2016: 7) and thus eliminates the need for the multitude of pre-processing measures and

steps that are involved in building a model for SMT. This chapter aims to give a concise and comprehensible explanation of the concepts and principles behind NMT and provides information on its current state.

First of all, the basic terms and concepts as well as their relations to one another shall be established. Fig. 12 is a very simple visualisation of the relations between the concepts Artificial Intelligence (AI), machine learning, and deep learning.

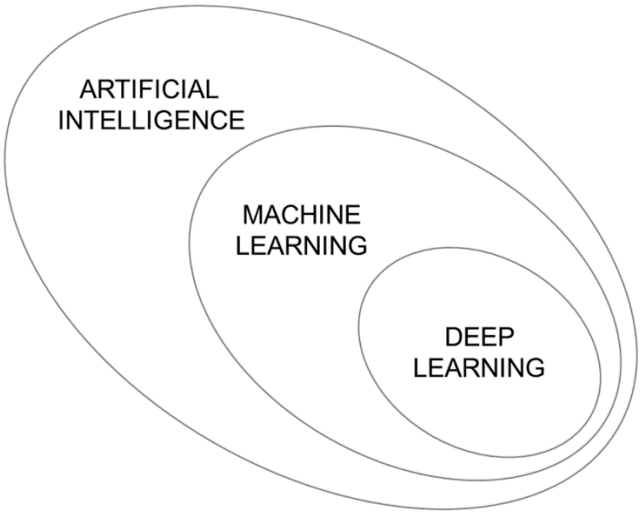


Figure 12: The relationship between artificial intelligence, machine learning, and deep learning (Kelleher 2019: 6)

The field which is concerned with creating the artificial neural networks that are used for NMT and that can process large amounts of complex data is known as deep learning (Kelleher 2019: 1). As can be seen in Fig. 12, deep learning forms part of machine learning, a domain that mainly deals with two topics: statistical machine learning and research in the area of neural networks (Kelleher 2019: 6). Machine learning, in turn, is a component of the wide and extensive field of artificial intelligence, which strives to create machines and computer programmes that are intelligent, i.e., that have the ability "to function appropriately and with foresight in [their] environment" (Nilsson 2010: xiii). Apart from Natural Language Processing (NLP), other areas and fields in which AI has been used or will be used in the future encompass homes (thermostats, kitchen appliances, doorbell cameras, vacuum robots, …), cars (navigation, intelligent parking systems, assisted or even autonomous driving, …), entertainment (gaming, virtual reality, book or film recommendations, …), medicine (devices for the prediction of heart attacks, systems which help to diagnose patients, …), business (rule management systems, scheduling software, …), and many more (Nilsson 2010: 615-629).

Although the inspiration for neural networks is the biological functionality of the human brain with its ability to form ideas and construct concepts by means of neural transmission (Poibeau 2017: 181), the relation between NMT and the actual brain is merely a vague one. A vast number of artificial units that are connected to each other make up the artificial neural networks that form the basis of NMT. These artificial units bear some resemblance to neurons in the way they work. The activation of the neural units/neurons, i.e., the extent to which they are stimulated in the sense of being excited or inhibited, is subject to the impulse the other neural units/neurons transmit to them and also depends on the weight that is assigned to the connections that are used for said transmission (Forcada 2017: 292). The value that these weights assume can be a positive or negative one; it provides information about "the strength and nature of [the neurons'] connection" (Forcada 2017: 293). This means, for instance, that if a connection that carries a positive weight is used to transmit an impulse to a neuron, said neuron in a state of excitement will be more likely to put the next neuron in the connection into an excited state as well. However, if the connection that is used for the stimulus has a negative weight assigned to it, an excited neuron will be more likely to put other connected neurons into a state of inhibition rather than excitement (Forcada 2017: 293).

This weighted input is only one part of the whole process. The other part is the actual activation of a neuron, which is usually confined to certain value ranges, e.g. between 0 and 1, between -1 and 1, or it can be limited to a positive value range only (Forcada 2017: 293). By means of what is known as an activation function, values of the weighted input can be mapped onto values of the activation. Depending on which type of activation function is used, this creates different sets of rules or conditions that control the relation between the two values, e.g. rules stating that the activation equals the weighted input if the weighted input is positive, or that the activation is 0 if the weighted input is negative (Forcada 2017: 293). Studying the activation of a single neural unit alone, however, is not very indicative yet. As the term 'neural network' implies, it is the interaction between the activations of many interconnected neurons that has an impact and leads to actual results.

ANNs consist of a vast number of layers. These layers can be seen as particular groups of hundreds of neurons or neural units, all of which are linked to the units in the next layer by weights. This way, thousands of connections are created (Forcada 2017: 295). From the activation states of the individual neural units within these layers, so-called distributed representations are constructed, which are considered to be a depiction of each neural unit's

state of activation (Forcada 2017: 294) and are also referred to as vectors. In a regular three-dimensional space, vectors have three real-valued numbers, i.e., coordinates, which can be used to determine the location of any point within said space. However, since language is a very complex concept with many facets and nuances, many more than three dimensions are required in order to appropriately represent words, sentences and the relations between them. Forcada speaks of "hundreds of [dimensions]" (2017: 295). Therefore, the vectors in this context consist of a multitude of numbers that are used to locate the respective words or sub-word units, e.g. characters or character sequences, in this multidimensional space that is utilised to represent language in vector space. The underlying idea is that words or sentences that belong to similar contexts or that represent similar ideas are located at a similar position, i.e., their vectors are in close proximity to each other in vector space, while vectors of words or sentences that are unrelated to each other or that represent very different ideas are located farther apart (Forcada 2017: 295).

The first step in the actual NMT process is, of course, training with large bilingual corpora. The training phase is a very resource-intensive endeavour, considering that it demands immense computational power and may take days or even months until it is completed (Forcada 2017: 295). The main goal of the training phase is for the output, i.e., target sentence, which the NMT system yields for each source sentence to be as close to the so-called gold standard as possible. The gold standard is the reference translation in the training set which, in the ideal case, has been produced by a professional human translator (Forcada 2017: 295). To accomplish this approximation to the gold standard, it is first necessary to determine how strong the links between the different neural units must be and which weights they need to have in order for the desired outcome to be produced. For this purpose, the training examples are processed, and during this procedure, the weights need to be updated in accordance with error terms (Koehn 2020: 77). The objective of these small weight corrections is to reduce the discrepancies between the output and the reference translation as much as possible, i.e., to strive for an error rate that approaches or virtually is zero (Forcada 2017: 296). After the training phase, there are different architectures or models that can be applied. Virtually all NMT systems have an encoder and a decoder, which is known as encoder-decoder model, and both the encoder and the decoder can be built with different neural architectures. For the encoder-decoder model and for all of the architectures that will be described, a so-called sentence embedding is required, i.e., one individual representation of the conveyed meaning of a given source sentence

(Koehn 2020: 199). The different architectures for NMT can use Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), or they can be equipped with a mechanism entitled (self-)attention, which is what the so-called Transformer model is based on. All of these terms and concepts will be explained in the following paragraphs.

The encoder-decoder model, in a nutshell, processes sequence after sequence. After the processing of a source sequence, e.g. a sentence, a unit of words, or an image, the compressed representation of the source sequence is forwarded to the decoder, which builds the target sequence, e.g. a sentence, a unit of words, or an image. This approach can function on the basis of any type of ANN, e.g. a Recurrent Neural Network (RNN), a type of ANN which is able to remember and store the network's output and subsequently feed it into the next step together with the next bit of input that follows in the sequence (Kelleher 2019: 170f.). In order to describe this procedure in more detail, to make it less abstract and easier to understand, Forcada (see 2017: 297f. for the entire process) uses the source sentence 'My flight is delayed.', which is to be translated into Spanish, as an example. At first, the word embeddings of the sentence's individual words learned by the model, e('my'), e('flight'), e('is'), e('delayed') and e('.') (e(…) = vector) are joined into one vector representation of the source sentence. Note again that each vector may consist of hundreds of real-valued numbers. Now the encoder can begin to perform the actual encoding process. Fig. 13 depicts this process and serves as a helpful visualisation.

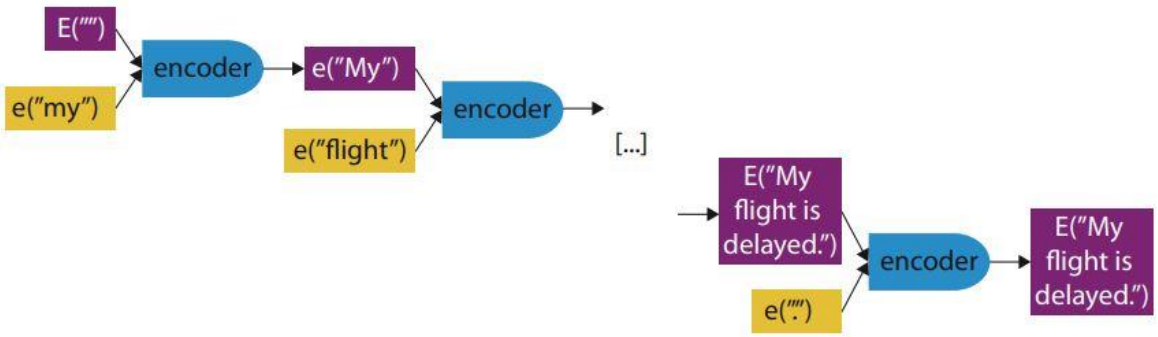


Figure 13: Encoding of the sentence 'My flight is delayed.' (Forcada 2017: 298)

Fig. 13 shows that in the first step of this procedure, the network which is responsible for the encoding process, the encoder, uses a pre-existing encoding, a type of 'shell' that can be seen as a placeholder for an empty sentence (represented as E('')) and couples it with the first word of the sentence (e('my')), which results in the encoding E('My'). This is where the

aforementioned functionality of the RNN comes into play; this very output encoding (E('My')) which the encoder has just created is immediately fed back into the very same network along with the next word embedding (e('flight')), which results in the encoding E('My flight'). This process is repeated until the entire source sentence is represented in this way (E('My flight is delayed.')). By means of this recurrence mechanism of RNNs, the network is able to draw on previous input in order to make 'better', i.e., more informed, decisions on how to proceed with the current bit of input (Kelleher 2019: 171). Nevertheless, current state-of-the-art NMT models fully rely on Transformers with attention mechanism that strongly outperform RNN-based architectures.

The second phase of the encoder-decoder model is, of course, the decoding process (see Forcada 2017: 298f. for the entire process). On the basis of the final encoding from the encoding process (E('My flight is delayed.')), the decoding network, or decoder, generates two different vectors. The first one is what Forcada calls an "initial decoder state D('My flight is delayed','), where '' represents an empty sequence of target words" (2017: 298). The second vector that the decoder generates contains the probabilities (p) for all words (x) which could potentially be suggested for the first spot of the target sequence: $p(x|$'My flight is delayed','). Several algorithms to optimise this calculation of probabilities and the filling of target sentence positions are standard in NMT systems, e.g. beam search, which assigns probabilities conditioned on the words and the probabilities of their previous positions. If the decoder has been trained well, it will consider the Spanish word 'Mi' to be the translation with the highest probabilistic value and output it. Like in the encoding process, the resulting output is fed back into the network for the next step. See Fig. 14 for a visualisation of the decoding process.

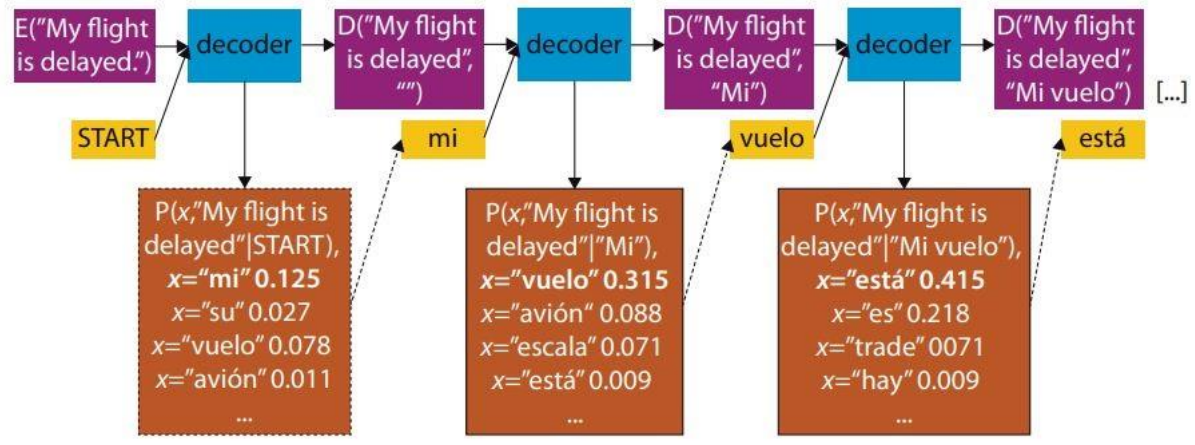


Figure 14: Decoding into Spanish of the representation of the English sentence 'My flight is delayed.' (Forcada 2017: 299)

As can be seen in Fig. 14, the decoding process continues in a similar fashion to the encoding process. The output of the first decoding (D('My flight is delayed','Mi')) and the vector containing the probabilities (p) for words (x) in the second position ($p(x|$'My flight is delayed','Mi')) are looped back into the process in order to obtain the next word in the target sentence. This procedure is repeated until the final target sentence has been formed in its entirety. The basic functionality of the encoder-decoder approach has thus been demonstrated with a rather simple example. Fairly quickly after the advent of this encoder-decoder sequence-to-sequence model, it was starting to be refined by means of a mechanism known as attention, which will be described in a later paragraph of this chapter.

An alternative architecture for NMT is the Convolutional Neural Network (CNN), which was originally designed to be used in the field of image recognition in the 1980s. While image recognition remains an area in which CNNs are commonly employed to this day, these networks have also been an important component of deep learning in the recent past (Koehn 2020: 198). The main advantage that CNNs bring with them is that they "reduce the dimensionality of input spaces" (Koehn 2020: 198). To put it in other words, they turn input representations with a high number of dimensions (for instance, in the case of NMT, this would be a vector) into increasingly smaller output representations that have a significantly lower number of dimensions by feeding them through several so-called convolutional layers. The following paragraph takes a closer look at the way this is achieved in the context of machine translation.

As is well known within the discipline of translation studies, the objective of (machine) translation in general should not be to produce a word-for-word translation in which each individual word is merely transferred from one language to another without taking context or implications into account. Instead, the aspiration should be to have the translated content express the same meanings, ideas, and nuances as the source material. As already mentioned earlier in this chapter, NMT thus requires a so-called sentence embedding of the source sentence. CNNs can be used in order to obtain such a single sentence embedding. Consider Fig. 15 for a better understanding of how exactly this process is carried out by means of CNNs.

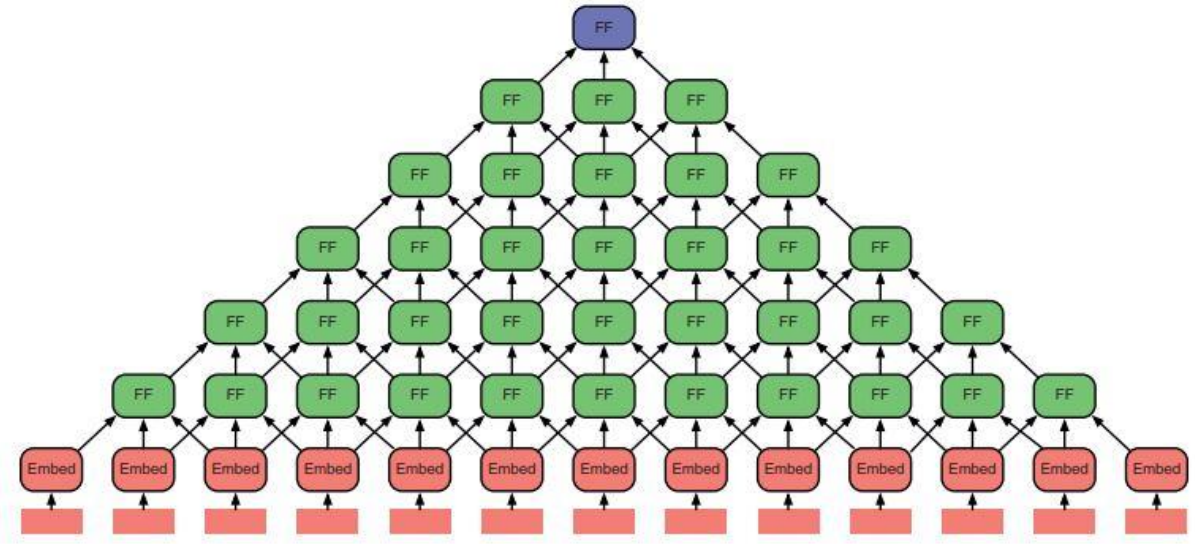


Figure 15: A convolutional neural network constructing a sentence embedding (top) from multiple word embeddings (bottom) (Koehn 2020: 199)

As is evident from the visualisation in Fig. 15, the process of computing a sentence embedding from multiple word embeddings takes place in several convolutional steps using so-called feed-forward layers. In each of these steps, at least three words (word vectors) are grouped together to create one new vector (Koehn 2020: 199), and inputs are usually overlapping, i.e., the second word of the first group of three is simultaneously the first word of the second group of three, the third word of the first group of three is also the second word of the second group of three, and so on. The grouping of the words and consequently the aforementioned overlap is displayed with arrows in each layer depicted in Fig. 15. With each layer that is passed through, the dimensions, i.e., the number of vectors, are reduced further and further until only one vector, the sentence embedding, remains. This is achieved by a number of convolution operations that, in contrast to RNNs, contain no cycles. The NMT system can then use this encoded single sentence embedding vector to decode it into a complete target sentence based on the representation of its meaning (Koehn 2020: 199). The type of architecture in the decoder is independent from the type of neural architecture in the encoder. In other words, a CNN encoder could also be paired with an RNN decoder.

A model that operates entirely without recurrence or convolution is the Transformer model using the concept of (self-)attention that was proposed by Vaswani et al. (2017). The attention mechanism was added to the encoder-decoder model almost instantly after the latter was developed. Expressed in simple terms, the attention mechanism causes the decoder to register and react to all previous representations, i.e., to the entire sequence of representations

that was created during the encoding process, not merely to the very last representation the encoder built (Forcada 2017: 299). As Koehn puts it, the attention mechanism "considers associations between every input word and any output word" (2020: 207) and thus also allows for refining the representations with words from the context while disregarding the distance between input and output (Vaswani et al. 2017: 2). Self-attention, in turn, is the concept of providing global dependencies between input and output (Vaswani et al. 2017: 2). The goal of self-attention is "to compute a representation of the sequence" (Vaswani et al. 2017: 2) without the need for recurrence or convolution. The reasoning behind the desire to rid NMT of recurrence and convolution operations is that due to memory constraints, both architectures become more and more critical the longer the sequences become (Vaswani et al. 2017: 2). Moreover, Vaswani et al. showed that the Transformer model, which is made up of encoder stacks and decoder stacks that, in turn, consist of multiple self-attention layers and sublayers, could reach new levels of quality of the translations despite requiring only a fraction of the computing power and training time in comparison to models using RNNs or CNNs (2017: 2). It is still the most widespread architecture utilised for NMT today.

Although NMT has already produced promising results in the past years, there are still cases in which the approach struggles to yield usable results, and areas in which there are some challenges to overcome (Koehn 2020: 293). These challenges can occur in a number of situations and include, among others, domain mismatch, insufficient training data, the translation of rare words, or the handling of so-called noisy data. The following paragraphs intend to provide an overview of the most common struggles that remain present in NMT, and of the drawbacks of using NMT in comparison to other MT approaches.

In terms of domain mismatch, it is difficult for the NMT system to find adequate translations if the text belongs to a different domain than the one the system was trained on (Koehn 2020: 294). This is due to the fact that one term might mean different things and represent different ideas and concepts in one field than it might in another. For instance, the noun 'administration' can be used as an example. If an NMT system was trained on parallel corpora containing texts from the field of law, it will most likely take 'administration' to mean the government of a country, or a group of people that is responsible for organisation and planning. However, in a text belonging to the domain of medicine, 'administration' will probably mean an entirely different thing, namely the act of giving medication to a patient. Consequently, if the NMT system trained with legal texts is then told to translate a medical text,

there is potential for error. Of course, this is only one example of a multitude of cases in which one word can have multiple meanings depending on the context. Koehn (2020: 294f.) conducted insightful research in this regard. He took five different MT systems, both statistical and neural, and trained each of them on corpora from different domains, namely law, medicine, IT, the Koran, and subtitles. In addition, a sixth system was trained using all available training data. The results of Koehn's study can be viewed in Fig. 16.

Source	*Schaue um dich herum.*
Reference	*Look around you.*
All	NMT: *Look around you.* SMT: *Look around you.*
Law	NMT: *Sughum gravecorn.* SMT: *In order to implement dich Schaue .*
Medical	NMT: *EMEA / MB / 049 / 01-EN-Final Work progamme for 2002* SMT: *Schaue by dich around .*
IT	NMT: *Switches to paused.* SMT: *To Schaue by itself . \t \t*
Koran	NMT: *Take heed of your own souls.* SMT: *And you see.*
Subtitles	NMT: *Look around you.* SMT: *Look around you .*

Figure 16: Translations of a sentence by systems trained on different corpora (Koehn 2020: 295)

Fig. 16 illustrates quite clearly that NMT still faces issues with regard to out-of-domain translation. While it performs well in terms of fluency and grammatical correctness, the translational equivalence suffers massively. Note, for instance, the translation of the system that was trained on the domain of the Koran; 'Take heed of your own souls' might be a perfectly fine sentence from a grammatical and stylistic perspective, but it has nothing to do with the intended meaning of the input sentence. The output from the NMT system that was trained on texts from the medical domain is even worse, seeing as it also contains numbers and acronyms, which were not present in the source sentence. The SMT output might not be perfect either, especially with regard to punctuation and untranslated words, but at least most of the SMT

output sentences contain correct words like ‘look’, ‘see’, or ‘around’, as was the intended meaning of the source sentence.

Another issue of NMT revolves around the amount of data used during the training phase of the system. Koehn (2020: 295f.) argues that although NMT has the potential to outperform SMT, it only does so after having been trained on a certain amount of data. In one of his studies, Koehn evaluated the performance of both NMT and SMT with training data volumes of different sizes using the Bilingual Evaluation Understudy (BLEU – for details on this evaluation method, see Chapter 2.3.4). He found that the learning curve of NMT tended to be a lot steeper than that of SMT, which was rather flat (Koehn 2020: 296). While this may initially sound like an argument in favour of NMT, the flipside was that NMT performed very poorly (BLEU score of 1.6) in the beginning stage, when the parallel training data was still at a volume of around 0.4 million words (Koehn 2020: 296). For comparison, SMT performed over ten times better (BLEU score of 16.4) at that stage with the same initial amount of training data. The point at which NMT gradually started to outperform SMT was not reached until around 15 million words had been used for training (Koehn 2020: 296). Summing up, it can be said that on the one hand, NMT can profit from high volumes of training data better than SMT can. On the other hand, however, NMT is a ‘slow starter’ in that it is unable to achieve reliable results with less than a few million words of training data (Koehn 2020: 296).

The translation of rare or even unknown words, i.e., words which hardly ever or never appear in the parallel corpora that are used in the training phase, poses another challenge for NMT systems. Due to the fact that these unknown or rare words tend to be named entities, i.e., proper nouns or numbers, a common approach is for the NMT system to simply copy the word to the target language without changes (Poibeau 2017: 192). In case the unknown word happens to be neither a proper noun nor a number, another option is to have the system break the unknown word down into sub-word units in order to “find relevant cues to help the translation process” (Poibeau 2017: 192). Simple sub-word unit approaches have been adapted to encode character sequences as Byte-Pair Encodings (BPEs) (Sennrich et al. 2016: 1717), which, even though proposed in the 1990s (see Shibata et al. 1999), more recently became the de-facto standard in language model training and machine translation. Contrary to the seemingly logical expectation that NMT systems would perform better if they encountered a rare word than if they encountered an entirely unknown one, Koehn made the surprising discovery that it is actually the other way round. In his research on the impact of infrequent words on translation

quality (see Koehn 2020: 297f. for details on this examination), he found that the NMT system he had trained translated unknown words correctly in 60.1% of cases. However, rare words, i.e., words which the system had encountered in the training phase, albeit only once, led to an even lower rate; in these cases, the system found the correct translation in only 52.2% of cases (Koehn 2020: 298).

The last problematic aspect of NMT that will be discussed in this chapter is the impact of noisy data on NMT. There is always the possibility that the parallel data used in the training phase might be noisy, i.e., 'polluted', as it were, with misalignments, wrong translations, content in a wrong language, e.g. untranslated bits of text, so-called non-linguistic characters, e.g. numbers, markup, bits of script language, etc., words that are in the wrong order, or other error types that interfere with fluency (Koehn 2020: 300). Koehn found that the presence of noisy data in the training corpora almost always affects NMT worse than SMT, and that NMT quality tends to decrease especially when one particular type of noise is prominent in the parallel data, namely the copying of source language segments into the target sequence (Koehn 2020: 299). In his experiment, Koehn artificially created this specific type of noise, namely "untranslated target sentence noise" (2020: 302) and added it to the training data of both NMT and SMT systems. The outcome was that the NMT system's performance plummeted from a 27.2 to a 3.2 BLEU score, while that of the SMT system was only reduced from 24.0 to 21.1 (Koehn 2020: 302). As a result of this exposure to untranslated target sentence noise, NMT systems will do what is known as 'overfitting the data', which means that the system will adapt to the noise to an extent that will render the system unable to "[generalise] to new data (data that is not in the training sample)" (Kelleher 2019: 20).

This chapter explained the basic concepts and workings behind neural machine translation and its components in a brief and easy-to-understand, yet comprehensive manner. It remains to be seen whether NMT will be able to outperform SMT constantly and completely or whether it might even manage to beat human translation in the future. After all, AI and deep learning are fields in which perpetual progress is being made. At the current point in time, however, it is fairly safe to assume that post-editing by a human will remain a necessary step in the NMT process in the foreseeable future.

2.2. Scientific translation

As is the case for the translation of most specialist texts, scientific translation involves a number of specific skills that are required in order to be able to produce an adequate and effective target text. Needless to say, it is not only crucial for (prospective) specialist translators to have excellent command of their respective working languages, which is the most important, yet maybe also the most obvious prerequisite, but they also need to be well prepared for and aware of various other aspects, such as domain-specific terminology, (inter-)cultural peculiarities, the communicative purpose and the intended audience of the source and target text, or genre conventions in both languages or target markets. This chapter describes the discipline of scientific translation and explains the aforementioned elements that form part of the process in more detail.

First, a delimitation between the terms 'science' and 'technology' shall be made. As Olohan (2016: 7) states, the relation between the two terms can be perceived and described in a multitude of ways. Sismondo defines 'science' as "a formal activity that creates and accumulates knowledge by directly confronting the natural world" (2010: 1) and 'technology' as "the relatively straightforward application of science" (2010: 7) or as "combin[ing] the scientific method with a practically minded creativity" (2010: 7). As can be derived from these definitions, science focusses more on gaining knowledge by observing and measuring, while technology is more concerned with putting the insights gained from science to use. In the context of translation studies, the two terms are sometimes grouped together for purposes of convenience (Olohan 2016: 7), in spite of the difference of meaning between them. Since the texts that have been used for the empirical part of this thesis (see Chapter 4.1.) belong to the scientific domain rather than the technical one, the rest of this chapter will focus on scientific translation, discourse, and language.

One aspect of specialist translation is that of domain-specific terminology and thematic knowledge. The majority of research publications are "written by specialists, for specialists, and […] non-specialist readers will feel excluded in the first instance" (Olohan 2016: 137). Consequently, this means that complex theories or ideas might not be discussed in detail because the writer presumes that the reader is already well versed in the concepts that form the basis for any new insights and contributions the publication will provide (Olohan 2016: 137). Therefore, it is vital for potential translators to be familiar with the subject, to possess profound knowledge about the respective topics and terms that are communicated in the text and to have an adequate range of subject-specific vocabulary at their command in both of their working languages. This can be achieved, for instance, by analysing and consulting parallel corpora and

using them for research regarding terminology and unfamiliar concepts (Olohan 2016: 138). However, even the most well-trained and well-prepared translators will most likely not end up having the same degree of proficiency as the scientists who write and read such publications (Olohan 2016: 137).

Translation is an "inherently social" (Olohan 2016: 15) process that not only requires language competence, but also intercultural competence. This means that it is necessary to be aware of country- or market-specific circumstances that might influence how a text or product is received in the respective locale. In the context of scientific translation, this could, for instance, apply to non-textual elements, i.e., any images, diagrams, or colours that a text might contain. What is perceived as logical and informative in one culture might be perceived as confusing or in some cases even inappropriate or offensive in another. Another aspect that may be affected by these intercultural differences is the structure of certain text types, or genre conventions in general. This topic will be discussed in more detail in a later paragraph of this chapter.

Another important element in the skill set of a translator of scientific content is the awareness of the text function, the communicative purpose and the intended audience of the source and target text. From the perspective of a translator, a text should always be considered a component of a communicative event (Olohan 2016: 6). Therefore, it is important to keep in mind why the text was written, what its desired function and communicative purpose is and whom it addresses. In the case of scholarly publications such as research articles, for instance, the sender and the recipient of the text usually belong to the same group of people, namely experts and specialists of the respective field. In most cases, the function of scientific texts is an informative one (Olohan 2007: 134), as is the case for research articles. Their communicative purpose is to share new knowledge and gained insights with peers from the same field. In other text types, however, the sender and the recipient may have different levels of specialist knowledge. Technical documentations or brochures, for example, are usually written by professionals and the intended audience are laypeople. While these kinds of texts are also considered to have an informative function, their communicative purpose is more of an instructive one, i.e., they intend to guide the reader through a certain process and give instructions. Books, on the other hand, might have a didactic communicative purpose, considering that they are often used in teaching settings, e.g. in university courses, or written for the general public.

Whatever the intended function and purpose of the source text may be, it is equally important to give thought to the translation commission, i.e., to what the purpose of the

translation is going to be. For instance, the translation might be targeted at a different audience, or it might not fulfil the same purpose as the original text. Maybe the translation is intended to be used for professional purposes, i.e., as a scientific publication, or maybe the commission only requests what is known as a gist translation, i.e., a mere summarisation of the basic source content in the target language. Gisting also plays an important role with regard to MT and is particularly relevant in scientific translation. It describes the act of specialists using unedited MT output in order to quickly grasp the contents of a foreign-language document in a very basic and general form (Nurminen 2019: 32). To sum up, many decisions in the translation process heavily depend on how and where the commissioner of a translation intends to use the target text.

In order for a text to fulfil its communicative purpose, it is necessary that said purpose also be accepted and respected by other people (Olohan 2016: 16). Especially if a specific kind of text or document is concerned, it is essential that the people affected by it be aware of the outcome of writing or reading this text. Olohan (2016: 16) uses a patent application as an example for this. Filing such an application would be utterly useless if the parties involved did not mutually acknowledge or agree on the consequences and impact that come with filing it. This shared understanding of a text's purpose is essentially what defines the notion of 'genre' (Olohan 2016: 16), which leads straight from the subject of the communicative purpose to the topic of text structures and different genre conventions in the field of science.

As far as translation in scientific discourse is concerned, it must be said that the genres that most significantly increase the translation demands within the field of science are research articles, e.g. in scientific journals, as well as their abstracts (Olohan 2016: 137), which is what the rest of this chapter focusses on. Other relevant genres to speak of include monographs, book chapters for edited volumes, reports, or conference proceedings. It is important to note that science as a discipline has changed over the past centuries, and therefore, so have the overall genre conventions of scientific research articles and papers. While science was once an almost recreational activity for men who could afford it, it has turned into a well-funded and profitable industry that occupies countless specialised experts all over the world (Olohan 2016: 149). Consequently, the way in which new scientific information and knowledge is disseminated is now more difficult to conceptualise. It is filled with a greater amount of information that is also becoming ever more complex, and it focusses more on the insights and results and less on the scientist who gained them (Olohan 2016: 149).

In anglophone contexts, the genre 'research article' usually consists of a number of different sections; an introduction, a chapter explaining the methods applied in the research process, a section presenting the results and findings, and another section dedicated to discussion (Olohan 2016: 150). Swales considers each of these sections to consist of 'moves', which he describes as "discoursal or rhetorical unit[s] that perform[s] a coherent communicative function in a written or spoken discourse" (2004: 228). A key word in this definition is "function"; as a functional unit, a move can be completed with as little as one phrase, but in other cases, it may as well be an entire paragraph containing several sentences (Swales 2004: 229). In other words, moves do not always strictly adhere to a certain formal structure. They can be divided into further smaller moves, called sub-moves or steps. See Fig. 17 for a list of moves that might make up the introduction of a research article.

Move 1: Establishing a territory	Making topic generalization(s) of increasing specificity (citations required)
Move 2: Establishing a niche	Step 1A: indicating a gap or Step 1B: adding to what is known Step 2: presenting positive justification (optional) (citations possible)
Move 3: Occupying the niche	Step 1: announcing present research Step 2: presenting research questions or hypotheses (optional) Step 3: definitional clarifications (optional) Step 4: summarizing methods (optional) Step 5: announcing principal outcomes (some fields) Step 6: stating value of present research (some fields) Step 7: outlining structure of paper (some fields)

Figure 17: Model of rhetorical moves and steps of introductions in research articles (Olohan 2016: 151)

According to the model in Fig. 17, which Olohan created on the basis of Swales' work (2004: 230, 232), most introductions of research articles tend to consist of three main moves, namely the establishment of a territory, the establishment of a niche within that territory, and the occupation of said niche. Each of these moves, in turn, consists of steps that serve a specific purpose. Such moves can also be identified for other sections, for instance for the discussion section, as can be seen in Fig. 18.

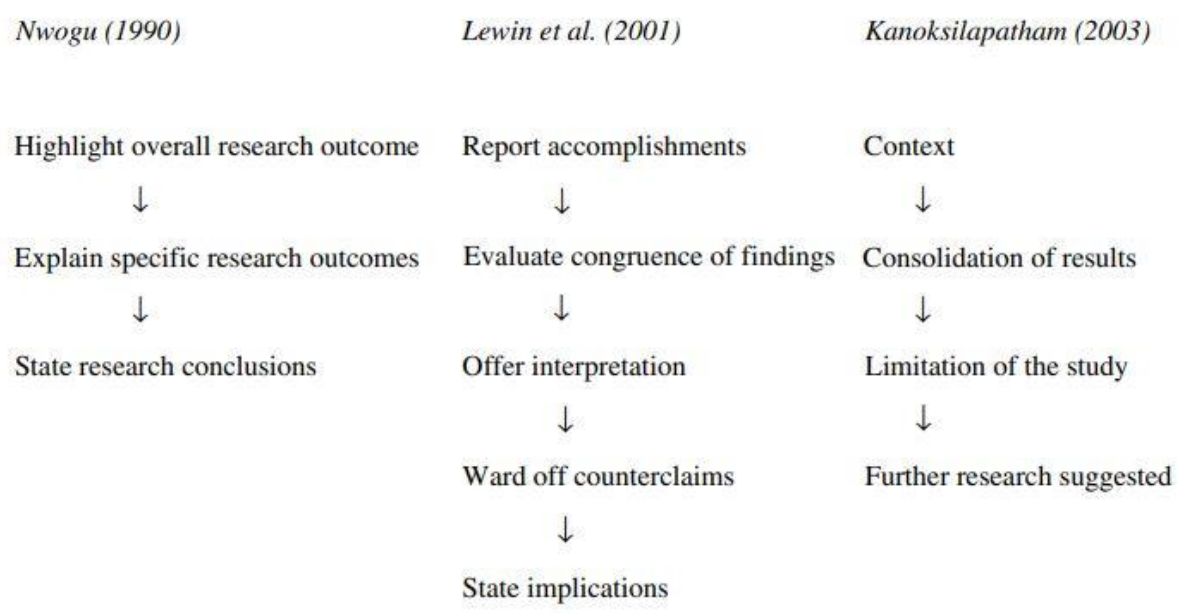


Figure 18: Three different structures with rhetorical moves for discussion sections (Swales 2004: 236)

Fig. 18 shows a comparison made by Swales of the way three different authors describe the structure and moves of a research article's discussion section. While Nwogu (1989)[1] only identifies three moves in a study of medical texts, Kanoksilapatham (2003) lists four moves based on texts from the field of biochemistry, and Lewin et al. (2001) provide even five different moves that can be applied in a discussion section for texts from the domain of social science. With regard to a research article's absolute number of moves, Kanoksilapatham (2003: 375) analysed 60 articles and was able to identify 15 moves in total, distributed across the four sections Introduction, Methods, Results, and Discussion. It should also be mentioned, however, that these models and findings mainly apply to "the monolingual anglophone context" (Olohan 2016: 155) and that they are likely to differ in other languages and locales. The structures and moves introduced in this chapter are merely intended to serve as a very basic example of what genre conventions might look like in the context of scientific translation.

The scientific language that is used within the domain of chemistry exhibits certain characteristics. Some of these characteristics are typical for academic writing and specialist language in general, such as impersonal style, unambiguity, the use of passive voice, the frequent occurrence of conjunctions and adverbs, and formal register, i.e., complete and grammatically correct sentences without the use of slang, colloquialisms or contractions. Other

[1] The official award and publication date of this thesis is 1989 and had been erroneously indicated as 1990 in Swales (2004: 236).

aspects, however, are specific to the field of chemistry, e.g. its domain-specific terminology, named entities, or its systematic nomenclature that was designed to not only allow for the naming of substances that were already known, but also for naming substances that might still be discovered in the future (Surman 2016: 134). Specialist terminology in chemistry may include names of elements, chemical compounds, certain procedures and concepts, descriptive terms that are used to define and explain the properties and behaviour of substances, or process verbs (Childs et al. 2015: 432), e.g. ‘distil’, ‘react’, ‘transfer’, ‘hydrolyse’, ‘separate’, ‘absorb’, to only name a few.

In summary, this chapter provided a concise overview of the skills a scientific translator requires and gave a glimpse into the kind of consideration that needs to go into the process of translating highly scientific texts. Moreover, the chapter explained the characteristics of scientific language, particularly in the domain of chemistry. Although the endeavour to translate content that is so highly specialised still remains a challenge even for professional translators, the procedure can be facilitated by developing a profound understanding of the purposes, functions, and characteristics that the genre brings with it (Olohan 2016: 149).

2.3. Quality of and evaluation methods for (machine) translation

Given the fact that the process of translation is a very complex one from various perspectives, including the cognitive, social, cultural, linguistic, and also technological one, it is no surprise that defining what makes a translation good and measuring the quality of a translation are endeavours that mirror this complexity (Castilho et al. 2018: 10). This chapter gives an overview of definitions and different interpretations of the term ‘quality’ both in human translation and in machine translation. Moreover, it lists and describes the most commonly used evaluation methods and metrics.

2.3.1. Definition of quality in translation

The definition and meaning of quality in translation usually depends on a number of factors. Different target audiences, industries, purposes, and contexts can influence how quality is perceived. Likewise, the actual quality standard of the end product also hinges on different circumstances, for instance on the perspicuity and unambiguity of the translation commission, on whether the source text is clean and free of errors, on the tools and resources (technology,

time, manpower) that are available, and many more. Defining what translation quality means and encompasses is crucial in order to be able to create evaluation methods that allow language service providers as well as researchers or students in the field of translation studies to objectively measure the quality of translation, be it human or machine translation (Koby et al. 2014: 415f.)

With regard to the aspects that distinguish a high-quality translation from a low-quality one, Gouadec (2010: 273) mainly lists criteria that apply to the target text only, such as fluency, efficiency, or readability. These aspects can be evaluated without also examining the original text, which means that the evaluation does not necessarily need to be carried out by a translator, but could, for example, also be performed by a monolingual person whose first language is the target language. However, Gouadec also mentions a criterion that can only be judged in comparison to the source text, namely the ergonomic degree of the translation, i.e., adequacy with regard to the enhancement of the original text and with regard to the adaptation of the content according to the cultural context of the target market (2010: 273). The implication that Translation Quality Assessment (TQA) should also take the source text into account is reinforced by Schäffner (1997: 1), who considers a high degree of accuracy to be the main feature of a good translation. Accuracy is a relational concept that describes "the correspondence between the source and target text" (Koby et al. 2014: 415), i.e., the characteristic of a translation being a "correct, precise, faithful, or true reproduction of the [source text]" (Schäffner 1997: 1) and thus also of the source text's message. In other words, it needs to be ensured that the intended purpose of the target text is considered during the translation process and thus reflected in the final product.

According to Koby et al. (2014: 415), however, accuracy and fluency are bound to be less than perfect to some extent, especially if the text wants to convey concepts or ideas that have cultural meaning or are in any other way specific to the source market. Instead of striving for flawless accuracy and fluency, Koby et al. propose to aim for the maximum degree of accuracy and fluency that is possible to attain within the boundaries of the respective translation process. For instance, there might be cases in which the party who commissioned the translation, the 'requester', might prefer a less accurate translation if it is truer to the formulation of the source text in exchange (Koby et al. 2014: 415). Another situation in which accuracy might have a lower priority is if the source text exhibits defects or errors. It might be desired by the

commissioning party that the defects also be recognisable in the translation, or the goal might be to conceal and improve the source text's defects (Koby et al. 2014: 415).

While the explanations and deliberations described in the two paragraphs above provide some sense of what it takes to produce a good translation, it would be beneficial and, frankly, more satisfying to be able to rely on an official, definite description of quality in translation. However, since researchers do not even necessarily agree on which activities the term 'translation' includes, it is clear that no such thing as a final and undisputed definition can exist for translation quality either. Nevertheless, Koby et al. (2014) have created two definitions, a broad one and a narrow one, that are quite concise, yet thorough.

The broad definition is more suitable for those who consider translation to not only include segment-by-segment translation, but also gisting, localisation, or transcreation, and who believe that general specifications cannot always be applied to any and all translation scenarios (2014: 416). This broad definition states the following:

> A quality translation demonstrates accuracy and fluency required for the audience and purpose and complies with all other specifications negotiated between the requester and provider, taking into account end-user needs. (Koby et al. 2014: 416).

This definition acknowledges that the target audience, the intended purpose, and the end user can all have or entail requirements, it states that most translations, by nature, have some extent of accuracy and fluency, and it addresses the fact that the requester and the provider of a translation must discuss and agree on the features and elements of the translation process and service (Koby et al. 2014: 416). The narrow definition, on the other hand, is more apt in the eyes of those who consider translation to be entirely text-based, i.e., who do not view the other aforementioned activities as translation, and who regard the majority of specifications as absolute (Koby et al. 2014: 416). This narrow definition reads as follows:

> A high-quality translation is one in which the message embodied in the source text is transferred completely into the target text, including denotation, connotation, nuance, and style, and the target text is written in the target language using correct grammar and word order, to produce a culturally appropriate text that, in most cases, reads as if originally written by a native speaker of the target language for readers in the target culture. (Koby et al. 2014: 416f.)

This narrow definition presupposes that many of the specifications are fixed in advance, and it involves a different range of use cases than the broad definition, seeing as the latter includes a much wider spectrum of activities. Consequently, deciding to stick with one of these

two definitions of quality will result in a different basis for the creation of evaluation and quality measurement methods (Koby et al. 2014: 417).

Now that the basic pillars of a high-quality human translation have been established, the aspects that can influence machine translation quality and the criteria for quality in machine translation can be examined briefly. In contrast to human translation, with MT quality there are other factors which contribute to how and to what extent the quality of the output can vary. The choice of the MT system type, e.g. rule-based, statistical or neural, as well as the size and domains of the bilingual corpora used in the training phase, the selected language pair and language direction, or the text genre are all aspects that can play a crucial role in influencing the results of MT (Schmalz 2019: 202). The criteria used to evaluate whether the MT is of good quality, on the other hand, are quite similar to the criteria outlined for human translation; accuracy, also sometimes referred to as fidelity, is not only a common indicator for human translation, but also plays an important role in MT evaluation, but also intelligibility (how easily the translation can be comprehended by a reader) or style (whether the language used in the translation is appropriate in relation to the purpose and content) are mentioned (Fiederer & O'Brien 2009: 54). The act of measuring and evaluating the quality of MT constitutes a vital part of advancing MT systems and thereby improving MT output.

2.3.2. TQA standards and models

As can be concluded from Chapter 2.3.1., it is already a challenging endeavour to specify what the criteria and definitions for quality in translation are. Bearing that in mind, it is unsurprising that Translation Quality Assessment (TQA) is an equally complex and broad-ranging topic, since it depends on several aspects and encompasses many methods. For instance, an evaluation of translation quality can have varying purposes and use different models that are in accordance with the environment in which it is carried out (Castilho et al. 2018: 11). If TQA is performed within the context of production or industry processes, the objective of the evaluation is to guarantee that the customer or end user is provided with a product that corresponds to a specified standard, i.e., a certain level of quality that needs to be met. If TQA is performed in an academic environment, e.g. for research purposes, the goal is usually to demonstrably measure, quantify and reveal any improvements that have been achieved in comparison to past translation approaches or previous research publications (Castilho et al. 2018: 11). This chapter

introduces some of the most common models, standards and metrics for assessing the quality of translation.

2.3.2.1. SAE J2450

In 2001, the Society of Automotive Engineers (SAE), more specifically General Motors, Ford, and Fiat Chrysler, created a quality metric to be used for translating service information, namely the SAE J2450. However, the metric only became a standard, i.e., a SAE norm, in 2005 (Martínez Mateo 2014: 78) after a revision by a committee consisting of language staff representatives of other member companies of the SAE. In their original draft from 2001, SAE stated that the aim of their metric was to create a standard that could be used as a reference against which it would be possible to measure the quality of a translation in the domain of service information in the automotive industry in an objective manner (SAE Mobilus 2001: 1). Moreover, they stated that their metric was intended to be applied without regard to the source or target language and regardless of whether the translation was produced by a human translator, with the help of a CAT tool, or by an MT system alone (2001: 1). The metric can be used by a human evaluator in order to identify errors and assign a penalty weight to these errors. Finally, a numeric score is calculated that specifies the translation's degree of quality. Fig. 19 provides a closer look at the error categories that are used in this metric.

Main Category	(abb.)	Sub-Category (abbreviation)	Weight serious-minor
Wrong Term	(WT)	serious (s)	5/2
Syntactic Error	(SF)	minor (m)	4/2
Omission	(OM)		4/2
Word Structure or Agreement Error	(SA)		4/2
Misspelling	(SP)		3/1
Punctuation Error	(PE)		2/1
Miscellaneous Error	(ME)		3/1

Figure 19: Error categories and sub-categories in the SAE J2450 norm (Martínez Mateo 2014: 79)

The seven different error categories used in the SAE J2450 norm depicted in Fig. 19, namely wrong term, syntactic error, omission, word structure or agreement error, misspelling,

punctuation error and miscellaneous error, can be assigned one of two sub-categories, namely serious (s) or minor (m), where each of these two sub-categories has its own weight. Additionally, the evaluator is provided with two meta-rules that are meant to help the evaluator with their decision in case of ambiguities. The first of these rules is to "always choose the earliest primary category" when in doubt (Martínez Mateo 2014: 79) and the second is to "always choose 'serious' over 'minor'" when in doubt (2014: 79). In order to determine the overall score for a document, the numeric weights of all errors that have been found across the seven categories are added up and the resulting number is divided by the number of source text words (Woyde 2001: 39). While this norm serves as a sound basis for reducing the types of errors that it includes and measures, it mainly focusses on linguistic and syntactic characteristics of language, whereas register or style are not considered. Consequently, this means that the metric is not ideal for evaluating texts in which style plays an important role. Moreover, Woyde also points out that "a translation that is free of any J2450 errors may still be unacceptable for other reasons" (2001: 39).

2.3.2.2. LISA QA Model

Another widely known standard was the LISA QA Model, which was created in the 1990s by the Localization Industry Standards Association (LISA) and was discontinued in 2011 when LISA terminated its activities entirely. Initially only available as a spreadsheet, the model was designed for the assessment of quality in translations from the domain of software localisation and documentation, and it came to be available as a stand-alone software later on (Lommel 2018: 112). The model functioned on the basis of error types that could be classified in three degrees of severity, namely minor, major and critical. Due to discrepancies between the model's documentation and its interface, it cannot be said with certainty whether the model consisted of 18 or 21 error categories (Lommel 2018: 112). The score of the evaluation hinges on the number of errors that were found, as well as on their types, and depending on the threshold value the evaluator has specified, the entire evaluation task either results in a 'pass' or 'fail' status (Castilho et al. 2018: 15). This model has been customised and adjusted several times throughout the years, for instance to fit other domains or text types. While some argue that this sort of TQA model has the advantage of providing standardisation to some extent, others consider this model's "underlying 'one-size-fits-all' philosophy" (Castilho et al. 2018: 15) to be too restricting.

2.3.2.3. Multidimensional Quality Metrics (MQM)

One of the many systems that are based on the LISA QA Model is the Multidimensional Quality Metrics (MQM) framework. As the name suggests, this framework is not one single metric, "but rather [defines] a common vocabulary to declare metrics" (Lommel 2018: 113). It is possible to use this model to evaluate both human translations and machine translation outputs, and to adjust the framework according to the user's preferences and needs (Castilho et al. 2018: 16f.). This means that the user can choose from a flexible catalogue of error types depending on the type of content they are dealing with. It is thus possible to determine and specify a quality metric that is based on the textual properties and characteristics of the content. MQM has hierarchies with up to four layers of depth, where each layer signifies an increasing degree of specificity (Lommel 2018: 116). The top level of the MQM hierarchy consists of eight different main branches, which are accuracy, design, fluency, internationalisation, locale convention, style, terminology, and verity (see Lommel 2018: 117f. for more detailed information on these branches). Each of these branches, in turn, can contain various issues on their lower and thus more specific layers. An example of the hierarchical structure of one of these branches can be viewed in Fig. 20.

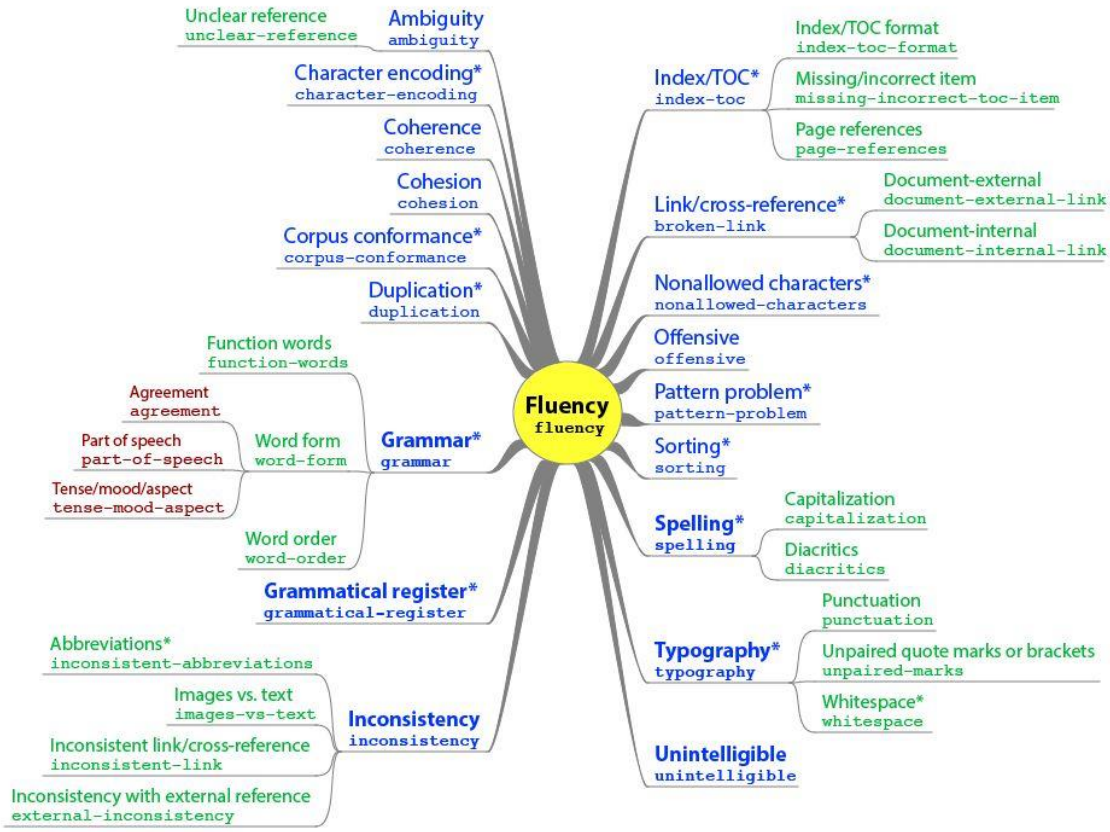


Figure 20: Hierarchical tree structure of the fluency branch of MQM (Lommel et al. 2015: n.p.)

Fig. 20 shows the 39 issues that are contained in the fluency branch in a tree structure, separated into different layers of increasing specificity. It is noted that there are always two designations for each issue; the top one is a human-readable name, while the bottom one serves as an identifier in programming languages (Lommel 2018: 116).

In order to create a metric from the MQM framework, it must first be decided which types of issues are capable of measuring and checking compliance with the specifications for the respective translation project (Lommel 2018: 120). During the assessment, evaluators have to assign a level of severity and importance to each error they encounter. The higher the severity, the higher the penalty points. If the default weighting is used, then null = 0, minor = 1, major = 10, critical = 100 penalty points. The higher the importance, the higher the weight (Lommel 2018: 121). The weight that is set per default is 1.0. Therefore, if an evaluator deems a certain branch or issue, e.g. fluency, to be very important for the particular text at hand, they can assign a higher weight to fluency, for instance 1.5. This would mean that all fluency errors count 50% more in terms of the penalty points that will be assigned to them. On the other hand, if the evaluator deems another branch or issue, e.g. design, to be less important for the respective

text, they can assign a lower weight such as 0.5 to design, resulting in a reduction of penalty points by 50% for all design errors that are encountered (Lommel 2018: 121).

The first step for determining the score of a translation evaluation is to calculate the penalty points. This is achieved by multiplying each error by its assigned severity value and its weight. The overall score is then obtained by dividing the penalty points by the word count of the translated text and subtracting the resulting quotient from 1 (Lommel 2018: 122). The resulting number lies between 0 and 1, and the score is then presented as the percentage that number represents. The scoring process will now be reviewed once more, this time with a concrete example. Sticking with the exemplary importance values from the previous paragraph, namely the weight of 1.5 for fluency errors, the weight of 0.5 for design errors, and the default weight of 1.0 for all other error categories, in this example it is assumed that in a translation with 1,000 words, an evaluator has come across one minor and one major fluency error, one critical accuracy error, and two minor design errors. Table 2 visualises these parameters and the resulting calculation of the penalty points.

Table 2: Calculation of penalty points in an MQM evaluation

		Weight	**Null (0)**	**Minor (1)**	**Major (10)**	**Critical (100)**	**Calculated points**
Error categories	Fluency	1.5		1	1		16.5 (1.5+15)
	Accuracy	1.0				1	100
	Design	0.5		2			1
						Total	**117.5**

Reiterating what was stated above, the weight needs to be multiplied by the respective severity value for each single error. As can be seen in Table 2, the two fluency errors with different severities thus result in 16.5 penalty points, the single critical accuracy error in 100 penalty points, and the two minor design errors in 1 penalty point. In total, this adds up to 117.5 penalty points, which must then be divided by the translation's word count (1,000), amounting to a value of 0.1175. In the last step, this value must be subtracted from 1, leading to a final evaluation score of 0.8825, or 88.25%.

People who apply the MQM framework must also define thresholds to specify which scores should be considered acceptable and which ones should not. One possibility to define

these thresholds is to ask the requester of the evaluation for reference translations, i.e., translations they deemed to be acceptable as well as ones they were not satisfied with, and to use the scores of these translations as a benchmark (Lommel 2018: 122). As mentioned before, the advantage of the MQM framework lies in the fact that it can be adjusted to fit the desired purpose. According to Lommel (2018: 122), this also enables users to selectively determine which factors cause qualitative shortcomings and which areas need improvement.

2.3.2.4. Karlsruhe comprehensibility concept

On the basis of the Hamburg comprehensibility concept, a model for the clear and comprehensible wording and formulation of specialist texts that was developed by Langer et al. (1974) and was of particular didactical usefulness in text production and optimisation courses (Göpferich 2009: 33), Göpferich introduced an extension of said concept in 2001, namely the so-called Karlsruhe comprehensibility concept. She claims that this concept is helpful in "optimising [non-instructive texts] with regard to features which quite obviously are detrimental to the text's comprehensibility" (Göpferich 2009: 32). In addition to the fulfilment of the communicative function, i.e., the extent to which a text complies with the specifications according to which it was composed, Göpferich (2009: 34) also considers the sender and the target audience as factors that may contribute to the overall text quality.

The originally four dimensions for features of text comprehensibility, which are (linguistic) simplicity, arrangement and (cognitive) structure, concision, and motivation, were extended by Göpferich by two more dimensions, namely correctness and perceptibility (2009: 34). She justifies her addition to the original concept by stating that her two new dimensions play an important part of text comprehensibility and are also decisive for further processing of the text (2009: 40), e.g. translation or TQA. After all, the evaluation of certain aspects of TQA, e.g. the adequacy with regard to the text's purpose, presupposes that the source text must be comprehensible.

2.3.3. Human evaluation of machine translation

Machine translation quality may be assessed by humans, however, there are some drawbacks to human evaluation. Humans are bound to lose focus after a certain amount of time or after having already examined a certain text volume, which may result in inconsistent evaluations

depending on the concentration of the evaluator. A possible solution for this problem is to measure the reliability of the scores. For instance, this can be achieved by having the evaluator assess the same translation at different times (Doherty 2017: 139). Another aspect that makes human evaluation difficult is subjectivity. If five different translators are asked to translate the same sentence, it is more than likely that the five target sentences differ from each other at least to some extent. This discrepancy can therefore also occur in assessment procedures, meaning that the same translation is assigned different scores or evaluation results by different evaluators. This issue should be tackled by determining the so-called inter-rater reliability, which examines how closely the scorings performed by each evaluator match each other for each translation (Doherty 2017: 139). Despite these pitfalls, there are several ways in which MT assessment can be carried out manually by human evaluators, among them error classifications, ranking, and test suites.

One of the methods is that of error analysis and error classification, a method which, if performed manually by a human, requires a high amount of time and resources. Although this method can also be (semi-)automated, Popović (2018: 129) argues that most tools are not capable of differentiating between error classes in detail. In most cases, error analysis aims to create an error profile that shows the "distribution of errors over the defined error classes" (Popović 2018: 131) and thus gives information about aspects of the respective tasks and facilitates decision-making processes, for instance with regard to the purchase, development, or use of an MT system. While some error classification models have already been discussed in Chapter 2.3.2., e.g. the MQM framework, there are still many others that feature different error categories. Vilar et al. (2006: 699) created an error classification depicted in Fig. 21 that was mainly intended to be used for SMT evaluation and that features noticeably fewer and less detailed error types than, for instance, the MQM framework. Another similar typology was proposed by Kirchhoff et al. (2012: 122), whose classification is based on Vilar et al. (2006) and consists of the categories 'Untranslated word', 'Missing word', 'Word sense error', 'Morphology', 'Word order error', 'Spelling', 'Superfluous word', 'Diacritics', 'Punctuation', Capitalisation', 'Pragmatic/cultural error', and 'Other'. This study examined user preferences of different error classes and the result was that the error types which were least preferred by the evaluators were 'Word order' and 'Word sense errors' (Popović 2018: 134). Another error classification that is worth mentioning and that features more error types is that of Stymne and Ahrenberg (2012) shown in Fig. 22, who examined the inter-rater agreement in their study and

found that simpler typologies result in a noticeably higher inter-rater agreement than error schemes that are very detailed (Stymne & Ahrenberg 2012: 1787).

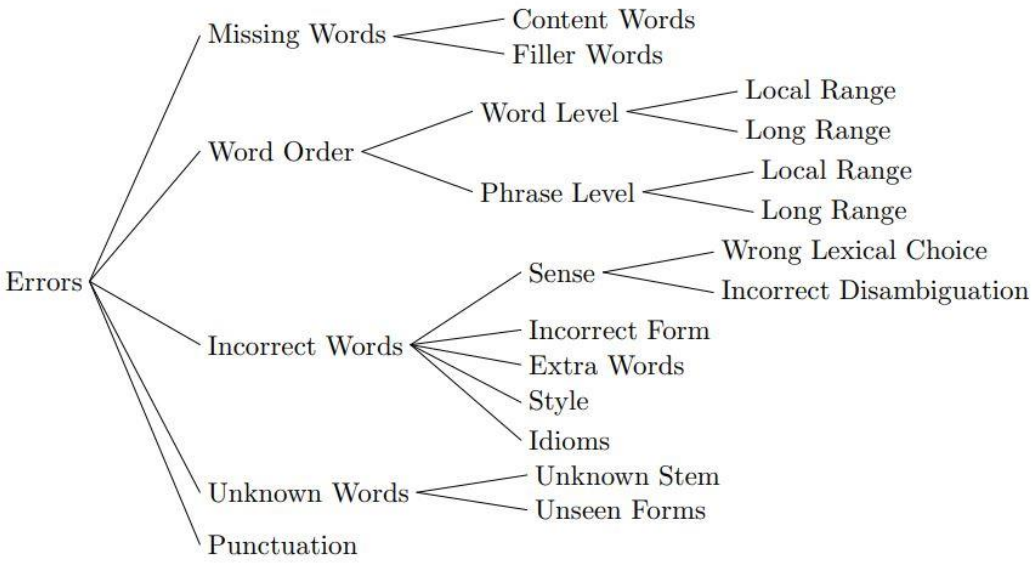


Figure 21: Error classification for SMT (Vilar et al. 2006: 699)

Label	Description
ER	Error-rate based categories: Missing, Extra, Wrong and Word Order
Ling	Linguistic categories, such as orthographic, semantic (sense) and syntactic
GF	Distinction between grammatical and function words
Form	Subcategories for morphological categories
POS+	Classifications of which category the error concerns, mainly part-of-speech based, but also contains other categories such as punctuation
FA	Judgments of whether the error concerns fluency, adequacy, both or neither
Ser	Judgments of the seriousness of an error on a 4-step scale from insignificant (0) to serious (3)
Reo	Cause of reordering
Other	Other distinctions
Index	The position an error has in the sentence. For most errors the position is marked in the system output, but for errors such as Missing words, the position is marked in the source sentence.

Figure 22: Error classification for MT (Stymne & Ahrenberg 2012: 1787)

Summing up the topic of error classifications, it can be said that the definition and the level of detail of the classification strongly depends on the respective task, but also on the language combination (Popović 2018: 140). Furthermore, it can be said that while more error types and more detailed classifications provide a better understanding of the errors, it is a lot more cognitively demanding for the evaluators to work with a plethora of error types and definitions and moreover, it results in lower consistency and reliability (Popović 2018: 154).

Another method to manually evaluate translation quality that is particularly popular in research contexts consists in ranking target texts. In this approach, a unit, e.g. a sentence, of the source text is available for reference, and several translations of said unit, each carried out by a different MT system, are provided in random order and without letting the evaluator(-s) know which text stems from which MT system (Castilho et al. 2018: 21). The evaluators are then provided with certain criteria or concepts according to which they are asked to arrange all the translations in order, namely from the best to the worst. This is the method that is used for the tasks in the Workshop in Machine Translation (WMT) series, not least because it is a comparably fast, easy, and efficient quality evaluation approach which yields results that can be interpreted easily (Castilho et al. 2018: 21). There are also other types of ranking, for instance one where the participants are asked to simply identify the best translation out of the selection options, or the so-called Best-Worst Scaling (BWS), in which the evaluators must decide on the best and the worst option out of each set of choices. Although BWS per se has existed for quite some time now (see Chapter 2.4. for more details on BWS), it is a rather innovative way to measure translation quality, which is why it was chosen for the empirical part of this master's thesis. So far, BWS has only rarely been applied in the context of TQA, e.g. in Finetti (2020) and Göschl (2022).

Another option for human/manual TQA that has not been used very widely so far is the concept of creating a test suite containing typical translation phenomena and errors as well as corresponding test sentences, which is what Burchardt et al. (2017) did with the aim of evaluating different rule-based, statistical (phrase-based), and neural MT systems in the language pair English-German. While this approach had already been used before Burchardt et al.'s study (2017) in the context of NLP for purposes of grammar checking, it had not been deemed useful for MT assessment up to the point of this study due to the difficulty of obtaining conclusions regarding the errors and the reasons for said errors (2017: 159f.). The test suite by Burchardt et al. (2017) contained about 800 entries for each language direction, comprising

"more than 100 different linguistic phenomena" (2017: 161). In the assessment process, the evaluators only focus on one phenomenon at a time. If the error in a sentence can be attributed to that phenomenon, the sentence is counted as incorrect. If, however, the sentence contains an error that is not related to the phenomenon at hand, the sentence is considered correct (2017: 162). This procedure is then performed for each of the phenomena in the test suite. The results can then be presented as percentages that indicate how good the different MT engines performed with regard to each phenomenon, respectively. According to Buchardt et al. (2017: 163), this evaluation approach does not deal with the degree to which a translation matches a reference, but rather, it examines "the nature of translations".

2.3.4. Automatic evaluation of machine translation

Of course, there are also a number of evaluation methods for MT that can be carried out automatically. Automated assessment methods offer the advantage that they are usually considerably faster than manual evaluation methods that need to be carried out by humans. Moreover, they are cheaper and the results they yield tend to be more consistent than those obtained via human TQA methods (Popović 2018: 130). This chapter covers some of the Automatic Evaluation Metrics (AEMs) that have been created and applied for machine translation: Word Error Rate (WER), Bilingual Evaluation Understudy (BLEU), and Translation Edit Rate (TER).

Based on the word level distance proposed by Levenshtein (1966) and originally used in the field of speech recognition, the Word Error Rate (WER) was first applied in MT evaluation by Nießen et al. (2000). This rate determines the word operations, i.e., the words that need to be inserted, deleted or substituted, which would be necessary to make the machine translation correspond exactly to a reference translation produced by a human (Castilho et al. 2018: 25). It does so by comparing the machine translation with the human reference translation and investigating the number of words in the target sentence that are firstly correct and secondly in the same position in both the MT and the reference (Way 2018: 165). Of course, this value must be standardised with regard to the respective reference sentence length. In other words, it is required to divide the number of erroneous words in the machine translation by the actual number of words in the reference sentence (Popović 2018: 142).

One of the most widely known and used automatic evaluation metrics is the Bilingual Evaluation Understudy (BLEU), created by Papineni et al. (2002). This evaluation method requires a numerical metric for translation closeness, which the creators based on the metric used by the WER, which works by means of *n*-gram matches, and a corpus of high-quality reference translations produced by human professional translators (Papineni et al. 2002: 311). The BLEU metric compares machine translations to the reference translations and determines how many *n*-gram matches they share, i.e., how many phrases consisting of *n* words are the same in both translations. Based on the outcome of this *n*-gram comparison, the system can then rank the machine translations. The MT output that has the highest *n*-gram agreement with the reference translation is classified as the best one and thus receives a higher score (Papineni et al. 2002: 312). On the other hand, words that appear in the machine translation but not in any of the reference translations receive penalties. Originally, the BLEU scores ranged from 0 to 1, with 1 being the closest to the human reference translation, but today, the score is commonly multiplied by 100 in order to facilitate the indication of progress (Way 2018: 163). BLEU has been found to exhibit high correlation with human judgment, particularly in comparison to the evaluation performed by monolinguals (Papineni et al. 2002: 318).

However, there are also downsides and limitations to the BLEU metric. Callison-Burch et al. (2006: 249) claim that a higher BLEU score does not necessarily mean an improvement of translation quality. BLEU does not lay down any rules for the order of corresponding *n*-grams, which results in the fact that a multitude of variations are allowed with regard to word choice (Way 2018: 168). Moreover, Callison-Burch et al. (2006: 251) argue that due to the fact that several reference translations are used in BLEU, there is a large number of translation variants that all end up with the same score, and that it is not to be expected that all these variants would receive identical scores if they were judged by human evaluators. BLEU also disregards what is known as recalls, i.e., "the proportion of the matched *n*-grams out of the total number of *n*-grams in the reference translation" (Banerjee & Lavie 2005: 67), which is a vital component in the evaluation of machine translation. Way (2018: 168) also claims that BLEU does not take semantic errors into account, which might result in mistranslations that still receive a high score. Nevertheless, Callison-Burch et al. (2006: 255) conclude that BLEU offers strong advantages in comparison to human evaluation, which is considerably more time- and resource-intensive, and that the justification for BLEU depends on the purpose for which it is used. According to them, the application of BLEU is suitable for "tracking broad, incremental

changes to a single system [or] comparing systems which employ similar translation strategies" (Callison-Burch et al. 2006: 255). They deem BLEU unsuitable for the comparison of MT systems that rely on entirely divergent concepts, e.g. phrase-based SMT and RBMT (2006: 255).

In order to overcome these limitations of BLEU, Banerjee and Lavie (2005) created the Metric for Evaluation of Translation with Explicit Ordering (METEOR). This approach attempts to perform better in terms of semantics and may even identify the same verb in different grammatical forms, e.g. "walking" vs. "walk", if semantic resources are implemented in the system (Poibeau 2017: 206f.). METEOR works by matching the MT output against the reference translation on a word-to-word basis and thus determining a score. In case of multiple references, the MT output is compared against each reference individually, and in the end, the highest value is documented (Banerjee & Lavie 2005: 67). While METEOR reports a higher correlation with human translation assessment than BLEU, it is more difficult to use the metric and interpret the outcome since there are more linguistic resources that might have been used (Poibeau 2017: 207). Despite its advantages, METEOR is therefore ultimately not applied as often as the BLEU score.

The last automatic evaluation method discussed in this chapter is the Translation Edit Rate (TER; also often referred to as 'translation error rate'), created by Snover et al. (2006). The basic functionality of TER is quite simple in that the rate determines how many editing operations would have to be carried out by a human in order for the MT output to be the same as the respective reference translation (Snover et al. 2006: 223). These edits are also known as 'shifts'. While the correlation of the TER approach with human evaluation is quite high, TER has been reported to overrate the actual rate of errors (Snover et al. 2006: 230). The authors noted that these correlations were merely exhibited on the basis of individual sentences and that the correlation would still need to be examined for larger segments (2006: 230f.). Nevertheless, the authors deemed TER to be "adequate for research purposes" (2006: 230).

2.4. Best-worst scaling

Best-Worst Scaling (BWS), the method which was chosen for the empirical part of this master's thesis, is a type of choice-based measurement method, and its basic premise is a rather simple one. Participants are shown multiple sets containing different options to choose from and are

then asked to select the two extremes among those choices, namely the best option and the worst one. In the context of MT evaluation, these different options would be different translations. This chapter takes a closer look at BWS and describes the underlying principles, the functionality, and the different ways in which this method can be applied.

2.4.1. History and underlying principles

The theoretical framework behind BWS is Random Utility Theory (RUT). It is the same framework that is applied in other types of choice experiments. In essence, RUT states that a person, when presented with different options, will choose the option that provides the maximum extent of utility to them. Simultaneously, RUT also acknowledges that random factors that cannot be anticipated or calculated, such as the person's preferences, attitudes, or perceptions, might affect the choice (Hess et al. 2018: 182). Subsequently, another underlying assumption of RUT is that the frequencies of a person's choices after having gone through the choosing process multiple times are a direct reflection of which of the provided options they value most (Louviere et al. 2015: 7).

While RUT had already been a topic of interest throughout the 20th century, BWS was first developed in the late 1980s by Louviere as a result of his curiosity regarding the potential benefits and insights that could be gained from determining participants' bottom choices in addition to their top choices (Louviere et al. 2015: 3). Louviere found that this additional piece of information allowed for a more comprehensive view of the utility function of any given respondent (Louviere et al. 2015: 3f.) and consequently, he spent the 1990s and early 2000s further developing the concept of BWS in collaboration with other researchers. Fields of BWS application included public polling in food safety, health economics, dental care, personality and values measurement, food and wine research, and many types of marketing research (Louviere et al. 2015: 4f.). Although Louviere et al. specify that it is possible to use BWS "in virtually all academic and commercial research applications in which category rating scales currently are used" (2015: xviii) and although BWS was proven to provide more dependable results than rating scales (Kiritchenko & Mohammad 2017: 469), there have, at the time of writing this thesis, only been few research papers (e.g. Finetti 2020, Göschl 2022) within the field of translation studies that have utilised BWS.

As already briefly mentioned above, the way BWS works is that participants are asked to select what they perceive to be the best option and what they perceive to be the worst option in a set. To put it differently, participants are required to identify the two options in the set that

are farthest apart from each other in terms of subjectivity (Louviere et al. 2015: xvii). Ergo, best-worst scaling is a human evaluation method, more specifically, as was already briefly discussed in Chapter 2.3.3., it counts as a type of ranking. Since this method is a "framework to measure latent, subjective quantities" (Louviere et al. 2015: xvii), the data received from the empirical research carried out in this thesis is a quantitative indication of how humans evaluate the quality of the translations.

2.4.2. Application of BWS

Louviere et al. (2015) describe different application cases of BWS, namely the BWS object case (Case 1), the BWS profile case (Case 2) and the BWS multi-profile case (Case 3). In order to make the differences between the three cases as comprehensible as possible, the examples for each case that are presented in this chapter all revolve around the same topic, namely around flights. Case 1, the BWS object case, is the simplest application. It merely lists a certain number of items, usually at least three, but ideally more, from which the participant must choose the best and the worst one (Louviere et al. 2015: 6f.). Fig. 23 gives an example of a BWS using the object case.

I think that this is the best Airline (☑ one)	Airline	I think that this is the worst Airline (☑ one)
☐	American	☐
☐	United	☑
☑	Qantas	☐
☐	Delta	☐

Figure 23: BWS object case (Louviere et al. 2015: 7)

The object case depicted in Fig. 23 uses different airlines as objects, where 'object' may refer to any list of items, and asks participants to choose which of the four listed airlines they think is the best and the worst one, respectively. Depending on the topic of the BWS, the terms 'best' and 'worst' may also be substituted for other superlatives such as 'most/least attractive', 'most/least adequate', 'most/least likely' etc. It is even possible to use entire phrases that constitute the two extremes on a subjective scale (Louviere et al. 2015: 6). For instance, if a municipality were to conduct a BWS with the intention of determining how much money to spend on which public policy issue, they might use phrases like 'I think we should spend the

most money on' and 'I think we should spend the least money on' as the extremes in their object case (see Louviere et al. 2015: 19).

In Case 2, the BWS profile case, the participants must select the most and least attractive attributes in a number of profiles that consist of several attributes each. Fig. 24 serves as an example for such a profile.

I think that this feature is the most attractive (☑ one)	Ticket option features	Specific details of Ticket Option	I think that this feature is the least attractive (☑ one)
☐	Airfare	$650	☐
☐	Travel time	6 hours	☑
☑	Airline	United Airlines	☐
☐	Freq flyer pts	Yes	☐
☐	No. of stops	None (direct)	☐
☐	Free drinks	No	☐

Figure 24: BWS profile case (Louviere et al. 2015: 9)

In the BWS profile case depicted in Fig. 24, Louviere et al. (2015: 9) list the attributes of a plane ticket, i.e., airfare, travel time, airline, frequent flyer points, number of stops and free drinks, as well as the values or levels of the respective attribute. The participants must decide which feature of this particular ticket, including its value, is most attractive and which feature, including its value, is least attractive in their opinion.

In Case 3, the BWS multi-profile case, participants are presented with sets of different profiles (Louviere et al. 2015: 9) and must choose the best and worst profile from that set. Fig. 25 is a visualisation of what a multi-profile case might look like.

Ticket option features	Ticket 1	Ticket 2	Ticket 3
Airfare	$650	$450	$550
Travel time	3 hours	5 hours	4 hours
Airline	United Airlines	Delta Airlines	American Airlines
Freq flyer pts	Yes	No	Yes
No. of stops	None (direct)	1	None (direct)
Free drinks	No	Yes	No
I'm most likely to choose (☑ one)	☑	☐	☐
I'm least likely to choose (☑ one)	☐	☐	☑

Figure 25: BWS multi-profile case (Louviere et al. 2015: 10)

As the name suggests, a BWS multi-profile case, such as the one depicted in Fig. 25, essentially features several profiles in one set, each of which has different values and levels for the same attributes. The participants must select one profile that they perceive as the most appealing and one that they perceive as the least appealing. In the case in Fig. 25, for instance, a younger person might select Ticket 2 as the best profile because it has the cheapest airfare, and because the person might not mind spending five hours on a plane. However, an elderly person might prefer Ticket 1, even though it is the most expensive one, because they might not physically be able to spend five hours sitting on a plane. Of the three application cases introduced in this chapter, the first one, namely the BWS object case, is the one that will be used for the research conducted in this thesis.

The most common way to go about creating the sets of options in a best-worst scaling is to construct so-called Balanced Incomplete Block Designs (BIBDs) (Louviere et al. 2015: 16). The concept of BIBDs is that none of the choice sets contain all available options, but rather that smaller subsets with different combinations of options are created. It must be ensured, however, that all options occur equally often across the entire BWS, i.e., that all options have the same number of chances to be picked as best/worst. BIBDs are particularly useful if there are so many options in total that featuring all of them in one single ranking might be overwhelming for participants. Although BIBDs can, in theory, be used for each of the three application cases, constructing them is a lot more complicated for Cases 2 and 3 than it is for Case 1. Moreover, the use of BIBDs for Cases 2 and 3 might lead to the situation that the number of levels of the same attribute might differ and that therefore, "many attribute-level combinations will not describe holistic options – that is, will not be profiles" (Louviere et al. 2015: 58). Fig. 26 and Fig. 27 exemplify the concept of a BIBD in a fictitious Case 1 BWS.

Object code	Object
1	Streets and roads
2	K–12 education
3	Tertiary education
4	Parks and recreation
5	Sports facilities
6	Housing developments
7	Job creation
8	Broadband access and speeds
9	Tourism facilities

Figure 26: Objects in a BWS object case (Louviere et al. 2015: 15)

Subset	Objects in each subset			Issues in each subset		
1	2	4	8	K-12 education	Parks and recreation	Broadband access/speed
2	1	4	5	Streets and roads	Parks and recreation	Sports facilities
3	4	7	9	Parks and recreation	Job creation	Tourism facilities
4	3	4	6	Tertiary education	Parks and recreation	Housing developments
5	1	2	3	Streets and roads	K-12 education	Tertiary education
6	2	5	7	K-12 education	Sports facilities	Job creation
7	2	6	9	K-12 education	Housing developments	Tourism facilities
8	1	8	9	Streets and roads	Broadband access/speed	Tourism facilities
9	5	6	8	Sports facilities	Housing developments	Broadband access/speed
10	3	7	8	Tertiary education	Job creation	Broadband access/speed
11	1	6	7	Streets and roads	Housing developments	Job creation
12	3	5	9	Tertiary education	Sports facilities	Tourism facilities

Figure 27: A BIBD for a BWS object case (Louviere et al. 2015: 17)

Fig. 26 shows a list of nine objects that shall be used in a Case 1 BWS as well as their respective object codes. The objects are public policy issues and the aim of this fictitious BWS is to determine how much money shall be spent on which of these issues. Fig. 27 shows what a BIBD for said nine objects might look like. This BIBD features 12 subsets, each of which includes three of the nine available objects. Each object occurs equally often throughout the entire BWS, namely four times, and each object co-occurs only once with each of the remaining objects.

Although Louviere et al. (2015: 18) provide a rather long but by no means exhaustive list of possible BIBDs (see Fig. 28 for a small excerpt of it), the formula for constructing a BIBD that is not on the list based on the parameters of one's own BWS is quite simple.

Design no.	Objects (v)	No. sets (b)	Occurs (r)	Set size (k)	Co-occurs (λ)
1	4	4	3	3	2
2	5	5	4	4	3
3	5	10	6	3	3
4	6	10	5	3	2
5	7	7	3	3	1
6	7	7	4	4	2
7	7	21	15	5	10
8	8	14	7	4	3
9	9	12	4	3	1
10	9	18	8	4	3

Figure 28: Excerpt of an illustrative list of potential BIBDs (Louviere et al. 2015: 18)

Fig. 28 shows a list of different BIBDs that are possible depending on the number of objects one wishes to include in a BWS. As can be derived, *v* represents the total number of available objects, *b* represents the number of sets in the BIBD, *r* represents the number of times each object occurs throughout the entire BWS, *k* represents the number of issues in each set, and λ represents the number of times each individual object co-occurs with each of the other objects. There are two underlying formulas that can be used to create individual BIBDs, namely $bk = vr$ and $\lambda(v\text{-}1) = r(k\text{-}1)$. However, since these parameters depend on and are tied to each other by these equations, it must be noted that, mathematically, not every desired combination of parameter values can be created (Louviere et al. 2015: 17).

Finally, this chapter will take a brief look at the way best-worst scalings are usually evaluated. The basic strategy is quite straightforward. One simply counts for each object how often it was picked as the best option and how often it was picked as the worst option throughout the entire BWS. In order to be able to rank the objects, however, it is useful to "[subtract] the worst count for each object from the best count for that object, and [to order] the objects by those scores" (Louviere et al. 2015: 21). This is where it proves advantageous to have gathered information not only about the participants' top choices but also about their bottom ones. For instance, if three out of nine objects were selected as the best option equally often, there would be no way of telling which of the three object is perceived as better or worse than the other two. However, if the number of times each of these three objects was selected as the worst option is counted as well, there is a greater chance of obtaining a clear ranking by subtracting the worst counts from the best counts (Louviere et al. 2015: 21).

3. Related work

This chapter gives an overview of previously published research articles dealing with the comparison of human translation against machine translation. It states the most essential parameters that constituted the research design of these research publications and summarises the insights gained from them.

Ahrenberg (2017) compared a human translation to the neural machine translation of the same text with Google Translate in the language direction English to Swedish. The aim of the research paper was "to assess the differences between state-of-the-art MT and HT" (2017: 21), to find out which characteristics define a machine translation in contrast to a human translation and to determine which actions a human translator can take that machine translation is not capable of (2017: 23). By analysing the number of shifted/unshifted clauses, where a shift is described as a translation procedure that does not fall under literal translation, e.g. paraphrasing, changing adjectival attributes into a relative clause etc., and by evaluating them using the BLEU and TER scores, Ahrenberg determined that MT did not yield results good enough for publication standards (2017: 25), that human translators not only produced grammatically intact sentences, but also stylistically better ones than MT (2017: 25) and that the human translator made use of methods that MT cannot yet apply (e.g. the shifting procedures mentioned above) (2017: 26).

Munková et al. (2019) compared human translation, machine translation by Google Translate, and post-edited machine translation in the language direction Slovak to German. Their aim was to compare quality and productivity of the translation of legal texts (2019: 239) and their method was to have 14 master students of translation studies, all of whom had Slovak as their first language, translate legal texts. Seven of these students produced a human translation and the other seven post-edited a machine translation. By using their own tool "wherein [they] implemented the metrics of automatic MT evaluation" (2019: 231), they checked the correctness of the translations and came to the following conclusions. Regarding productivity, the post-editing of machine translations proved to be more efficient than producing a human translation. Regarding quality, however, human translation achieved better results than machine translation, but similar results to the process of post-editing a machine translation (2019: 231).

Läubli et al. (2018) published research that compares neural machine translation to human translation in the language direction Chinese to English and additionally focusses on the contrast of document-level evaluation and the evaluation of isolated sentences. Their hypothesis was that a translator would perceive a human translation as better than a machine translation in

a document-level evaluation (2018: 4791) and that "professional human translation is still superior to neural machine translation" (2018: 4792). Läubli et al. asked professional translators with a minimum of three years of experience to pick the better translation in a set of two: one human translation and one machine translation. In other words, they used pairwise ranking. The raters were asked to make their decision based on the two criteria adequacy, where the raters were shown both the source text and target texts, and fluency, where the raters were shown the translations only. In random order, the translators rated entire documents and single sentences. The results showed that as far as adequacy is concerned, the HT and the MT do not show significant differences in isolated sentences (2018: 4793), but at the document level, HT was perceived as noticeably better than MT.

In their study, Popel et al. (2020) compared neural machine translation to human translation in the language direction English to Czech. A special focus was put on the preservation of text meaning, i.e., on translation adequacy. They presented and used their own neural translation system called Charles University Block-Backtranslation-Improved Transformer Translation (CUBBITT), which had previously outperformed the translation produced by a professional translation agency on the sentence level in the Workshop in Machine Translation (WMT) 2018 competition (2020: 2). Since that result from 2018 was perceived as surprising, Popel et al. conducted the research described in this publication, which was more challenging for MT as it also took context from the surrounding sentences into consideration. Professional translators as well as non-professionals were asked to annotate the translations with regard to errors. The annotators also had to specify if they only recognised the error due to the surrounding context (2020: 5). The researchers determined that CUBBITT did a better job than the human translator with regard to spelling, grammar and adequacy, while the human translation showed fewer fluency errors. Moreover, the neural translation system made more context-related errors.

Vasheghani Farahani (2020) conducted research comparing the results of the two different neural machine translation tools Google Translate and Microsoft Bing Translator with human translation in the language direction English to Persian with regard to the degree to which the translation matches the source language content (2020: 84). The research paper aimed to answer the question whether the output of a machine translation system was sufficient in its adequacy in order to compare to human translation (2020: 84). In the first step, Vasheghani Farahani instructed two professional and experienced translators to create translations of each of the seven source language texts to be used as a reference in the evaluation process later on. Then, he used the NMT tools Google Translate and Microsoft Bing Translator to obtain two

different NMT outputs, while two translation students, considered non-experts, were responsible for producing the human translations against which the NMT outputs would later be compared (2020: 89). The first student translated three of the seven texts, the other one translated the remaining four. After the evaluation of the outputs using a model for the prediction of translation adequacy proposed by Specia et al. (2011), Vasheghani Farahani found that while MT systems such as the two used in the study are indeed capable of creating a translation that exhibits sufficient adequacy to be considered equal to a human translation, the end product still requires human post-editing in order to make up for any other inaccuracies the MT systems might have caused (2020: 101).

While the five publications discussed in this chapter all had the same basic objective, namely to compare machine translation to human translation, they all put their focus on different elements and criteria of translation, different language pairs, as well as the usage of different tools and evaluation methods. In summary, it can be said that although there are individual cases in which, due to specific circumstances or aspects, MT can perform better than a professional human translator, the findings of these publications generally point to the assumption that MT, especially MT without any further human intervention such as post-editing, is not (yet) capable of yielding results that fully compare or even are superior to human translation.

4. Methodology

This chapter describes the process behind the empirical part of this thesis and outlines the research method that was chosen for attempting to answer the research questions (see Chapter 1.1.) that this master's thesis intends to examine. First, a short and concise overview of the entire undertaking is provided and then, the individual steps are described in more detail in the respective subchapters. The basic intention is to select a number of English-language source texts from the field of chemistry, to produce one human translation and four neural machine translations of each of those source texts from English into German and to subsequently conduct a BWS survey with the aim of rating the quality of the translations and thus simultaneously also rating the translation mode as well as the tool that was used to produce the respective translation. In addition to selecting the best and worst translation of each set, participants will also have to assign plus points from 0 to 4 to the best translation and minus points from -4 to 0 to the worst translation. The awarded points are then added up during the evaluation and each translation mode/tool thus receives a second numerical score in addition to the score that results from subtracting the worst counts from the best counts.

The empirical research method chosen for this master's thesis, i.e., for rating the translations and the tools/mode behind them is Best-Worst Scaling (BWS), which was already described in detail in Chapter 2.4. of this thesis. Best-worst scaling is a human evaluation method, which was deemed most useful for this master's thesis because it takes subjectivity of evaluation into account and because some automatic metrics have been found to exhibit an "inherent bias" against systems that do not work on an *n*-gram basis, such as NMT (Way 2018: 169). Of the three possible BWS cases (described in Chapter 2.4.2.), the first one, namely the BWS object case, is the one that is relevant for the research conducted in this thesis. The five objects in this BWS survey are human translation, DeepL Translator (2022), Google Translate (2022), Microsoft Bing Translator (2022) and SYSTRAN Translate (2022).

4.1. Text selection and target groups

The first step of the process consisted in selecting the source text passages, five in number with an approximate length of 30 to 50 words each. The selected passages are all excerpts from highly specialised texts, e.g. research articles or scientific publications, belonging to the domain of chemistry that were published in English language. Moreover, the different passages were

all taken from texts that deal with the same subject area within the domain of chemistry, namely the topic ‘dyes and pigments’. The publications were searched for and obtained through the library search engines of the University of Vienna and the Vienna University of Technology as well as through Google Scholar.

An important criterion for selecting the passages was that they must conform to the characteristics of the specialist and scientific language that is typical for texts from the domain of chemistry described in Chapter 2.2., i.e., they must exhibit features such as formal register (“could not” as opposed to “couldn’t”), impersonal style, unambiguity, the use of passive voice (“was eliminated”, “was attributed to”), or the frequent occurrence of conjunctions and adverbs (“and yet”, “in order to”, “with regard to”, “therefore”). Moreover, the text passages must include domain-specific terminology, e.g. named entities such as names of elements (“chrome”, “cobalt”), chemical compounds (“hydroxyquinolone”, “pyrazolone”), certain procedures and concepts (“application procedure”, “esterification reaction”, “levelling power”), descriptive terms that are used to define and explain the properties and behaviour of substances (“dehydrating”, “substantivity”, “wet- and lightfastness”), or process verbs (“charge”, “eliminate”, “hydrolyse”, “promote”, “bind with”). The selected source text passages were extracted from publications that deal with the topic ‘dyes and pigments’, namely from two different research articles that were published in scientific journals, and from a monograph on the same topic. The passages as well as their sources can be found in Appendix 1 of this thesis.

The participants of the survey must belong to one of the following two target groups. On the one hand, the survey addressed trained professional translators whose working languages include English as well as German. On the other hand, it addressed natural scientists, ideally chemists, who have good command of both English and German. Regardless of the target group they belong to, participants should have a high degree of expertise in their respective field since the texts are highly scientific and specialised. To be more precise, respondents should have at least a bachelor’s degree in their field. The goal was to receive at least six responses in each target group, i.e., at least 12 responses in total.

4.2. Translation process

The next step in the process was to produce the translations for the selected source text passages, that is, one human translation and four neural machine translations of each passage. For the

human translation, the author of this thesis opted to hire a professional translation agency that focusses on specialist translations, including the translation of chemical and pharmaceutical texts. There were two reasons for this decision. Firstly, chemistry is not one of the fields the author has specialised in, which is why the author is not too familiar with the respective terminology. Secondly, the author felt that hiring a neutral third party for this task would ensure an unbiased, more independent and thus more reliable translation.

The machine translations of the source text passages were produced using four different NMT tools, namely DeepL Translator, Google Translate, Microsoft Bing Translator and SYSTRAN Translate. These tools were selected because they are the most common and most widely known machine translation tools that are commercially available online and for free. All of them are very intuitive and easy to use, and they do not require any installation or preparatory steps other than navigating to the website and entering the text one wishes to translate. Fig. 29 shows the user interface of DeepL Translator as an example.

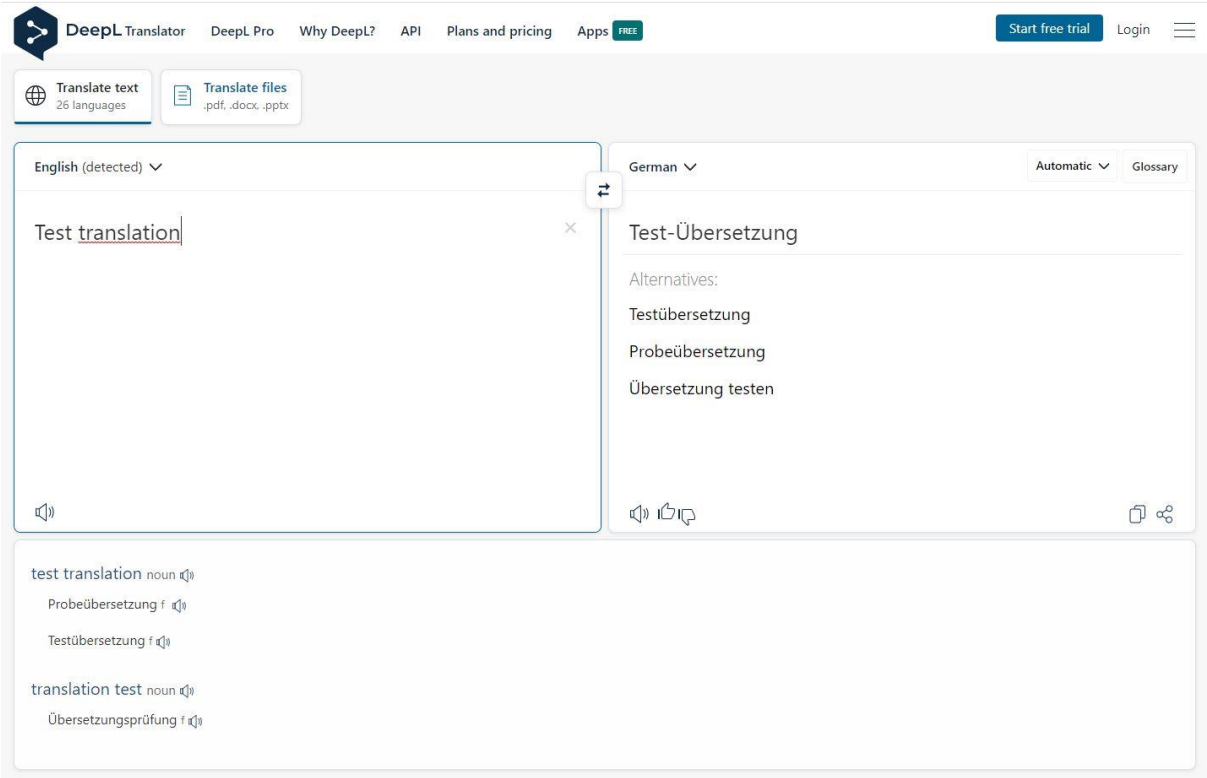


Figure 29: User interface of DeepL Translator (DeepL Translator 2022)

The basic user interface shown in Fig. 29 is also representative of the other three NMT tools used in this thesis. At the very top, the user can select whether they wish to translate text or entire files. Below, there are two large text fields with the option to select the desired

languages, which are 26 in total for DeepL Translator. The source language field is on the left side and the target language field on the right side. As soon as text is entered in the source language field, the tool starts to translate the words in the target language field. If the user is not entirely satisfied with the output, they can either manually select one of the alternative suggestions that are displayed below the output, or they can click on individual words of the translation and choose one of the synonyms suggested for it. The exception to this design is Microsoft Bing Translator, which does not offer the feature of suggested synonyms or alternative translations. In any case, however, this feature is not relevant for this survey since it is the raw, unedited output of the tools that is subject to comparison and evaluation. In other words, no choices regarding provided alternative translations or synonym replacements were used to produce the machine translation to be evaluated for any of the tools included.

Once all the translations had been created, it had to be ensured that the five different translations for each source text passage differed from each other to an extent that made a comparison and evaluation against each other viable and reasonable. Table 3 shows a comparison of the five translations for one of the text passages as an example. The differences between them are highlighted in red.

Table 3: Differences between the five translations

Source text	Selected acid 1:1 chrome complex dyes, because of their good wet- and lightfastness and their very good levelling power, and also 1:2 chrome complex and 1:2 cobalt complex dyes with hydrophilic groups, have successfully come into use as silk dyes.
Human translation	Erfolgreich als Seidenfarbstoffe eingesetzt werden ausgewählte saure 1:1-Chrom-Komplexfarbstoffe, aufgrund ihrer guten Feuchtigkeits- und Lichtbeständigkeit sowie der sehr guten nivellierenden Eigenschaften, sowie auch 1:2-Chrom- und 1:2-Cobalt-Komplexfarbstoffe mit hydrophilen Gruppen.
DeepL Translator	Ausgewählte saure 1:1-Chromkomplexfarbstoffe haben sich aufgrund ihrer guten Nass- und Lichtechtheit und ihres sehr guten Egalisiervermögens, aber auch 1:2-Chromkomplex- und 1:2-Kobaltkomplexfarbstoffe mit hydrophilen Gruppen als Seidenfarbstoffe bewährt.
Google Translate	Als Seidenfarbstoffe haben sich aufgrund ihrer guten Naß- und Lichtechtheit und ihres sehr guten Egalisiervermögens ausgewählte saure 1:1-Chromkomplexfarbstoffe sowie 1:2-Chromkomplex- und 1:2-Kobaltkomplexfarbstoffe mit hydrophilen Gruppen erfolgreich durchgesetzt.
Microsoft Bing Translator	Ausgewählte Säure-1:1-Chromkomplexfarbstoffe sind aufgrund ihrer guten Nass- und Lichtechtheit und ihrer sehr guten Nivellierleistung sowie 1:2-Chromkomplex- und 1:2-Kobaltkomplexfarbstoffe mit hydrophilen Gruppen erfolgreich als Seidenfarbstoffe in Den Einsatz gekommen.
SYSTRAN Translate	Ausgewählte saure 1:1-Chromkomplexfarbstoffe sind aufgrund ihrer guten Nass- und Lichtechtheit und ihres sehr guten Verlaufsvermögens sowie 1:2-Chromkomplexfarbstoffe und 1:2-Kobaltkomplexfarbstoffe mit hydrophilen Gruppen erfolgreich als Seidenfarbstoffe in Einsatz gekommen.

As can be seen in Table 3, the five translations for this particular source text passage differ from each other in some respects. Firstly, there are some orthographical differences, e.g. ‘Chrom-Komplexfarbstoffe’ vs. ‘Chromkomplexfarbstoffe’ or ‘Kobalt’ vs. ‘Cobalt’, but also actual orthographical errors that were made by the NMT tools, e.g. ‘in Den Einsatz’ or ‘Naß-’. Some variation in the area of domain-specific terminology can be found as well, for instance in the case of ‘levelling power’, which was translated as four different terms in German, namely ‘nivellierende Eigenschaften’, ‘Egalisiervermögen’, ‘Nivellierleistung’ and ‘Verlaufsvermögen’, or in the case of ‘wet- and lightfastness’, which was translated in two different ways in German, namely as ‘Feuchtigkeits- und Lichtbeständigkeit’ and ‘Nass- und Lichtechtheit’. Moreover, there are several differences that concern the sentence structure. While most of the NMT tools produced the translation in a manner that adheres closely to the source text’s structure, the human translator and Google Translate have shifted the sentence structure so that the translations start with the phrase that appears at the very end of the original sentence.

4.3. BIBD and arrangement of the target texts

Next, the balanced incomplete block design for the best-worst scaling was created. As discussed in Chapter 2.4.2., with the parameters that are known, it is possible to calculate the remaining parameters by means of two different equations, namely $bk = vr$ and $\lambda(v\text{-}1) = r(k\text{-}1)$. Two parameters were already known and fixed in this particular BWS, namely the total number of objects ($v = 5$, namely human translation, DeepL Translator, Google Translate, Microsoft Bing Translator and SYSTRAN Translate) and the number of issues in each set ($k = 4$). Based on these two variables, it would have been possible to do either five sets ($b = 5$), in which case each object would have had to occur four times ($r = 4$) and co-occur with each other object three times ($\lambda = 3$), or ten sets ($b = 10$), in which case each object would have had to occur eight times ($r = 8$) and co-occur with each other object six times ($\lambda = 6$).

Due to the fact that this is a human evaluation method and humans tend to lose focus after having already examined a certain text volume, it was decided that ten sets with four translations each, i.e., 40 translations in total, would be too many and would go beyond the scope of a survey of this scale, especially given the high degree of domain-specificity of the texts. The risk of participants abandoning the survey before completion was deemed too high. Therefore, the design featuring five sets with four translations each, i.e., 20 translations in total, was chosen. Table 4 visualises the structure of the design.

Table 4: BIBD for the BWS of this thesis

Set 1	H	G	D	S	5 objects, 5 sets, 4 items per set
Set 2	M	H	G	D	Each object occurs 4 times in total
Set 3	S	M	H	G	Each object co-occurs with each of the other objects 3 times
Set 4	D	S	M	H	Each set represents a different text passage
Set 5	G	D	S	M	
$v = 5$					Object D = DeepL Translator
$b = 5$					Object G = Google Translate
$r = 4$					Object H = human
$k = 4$					Object M = Microsoft Bing Translator
$\lambda = 3$					Object S = SYSTRAN Translate

Table 4 illustrates the parameters of the BWS and shows how the five different objects are distributed across the five sets in the BIBD. Each object has been assigned a different colour

and its own abbreviation. As the table shows, none of the sets are the same, i.e., no combination of four objects is repeated. Thus, each object has the same chances of being picked as best or worst by the participants. In the survey, each set features a different text passage. The full sets including the actual source text passages and their four respective translations can be found in Appendix 2.

In order to test the validity of the research method, the survey also tests the so-called intra-rater reliability. The objective of testing the intra-rater reliability is to examine the respondents' self-consistency (Gwet 2014: 340), i.e., to examine whether the respondents, when asked the same question at two different points in time, give the same answer on both occasions. For this purpose, one of the five sets is repeated exactly as it is, i.e., with the same source text and the same translations, only at a later point in the survey. Ideally, the repeated question should be inserted with some distance to the original question in order to avoid that the respondent has still memorised which options they picked as best and worst for the original question. Therefore, Set 2 is presented again after Set 5. Although it is completely identical to Set 2, it is renamed to Set 6. Table 5 shows the final arrangement of the texts and the individual number that has been assigned to each target text.

Table 5: Arrangement of the texts

Source text	Translation tool/mode	Target text
1	1. Human	1.1
	2. Google Translate	1.2
	3. DeepL	1.3
	4. SYSTRAN Translate	1.4
2	1. Microsoft Bing Translator	2.1
	2. Human	2.2
	3. Google Translate	2.3
	4. DeepL	2.4
3	1. SYSTRAN Translate	3.1
	2. Microsoft Bing Translator	3.2
	3. Human	3.3
	4. Google Translate	3.4
4	1. DeepL	4.1
	2. SYSTRAN Translate	4.2
	3. Microsoft Bing Translator	4.3
	4. Human	4.4
5	1. Google Translate	5.1
	2. DeepL	5.2
	3. SYSTRAN Translate	5.3
	4. Microsoft Bing Translator	5.4
6	1. Microsoft Bing Translator	6.1
	2. Human	6.2
	3. Google Translate	6.3
	4. DeepL	6.4

The cells highlighted in red in Table 5 indicate the source and target texts that are identical. The responses and plus/minus points obtained from Set 6 will not be counted in the evaluation since this would lead to an imbalance in the distribution of the five objects. In other words, if the responses from Set 6 were considered in the evaluation, four of the five objects would have a higher chance of being picked as best or worst, which would falsify the results of the survey. This procedure is merely applied in order to test the consistency of the responses of each individual participant.

4.4. Creation of the survey

The next step was to create the online survey. The tool LimeSurvey (LimeSurvey 2022), which is available online and for free, was chosen for this purpose. The participants' answers can be saved on the LimeSurvey page and exported into different formats, e.g. .xlsx or .csv.

Since it is reasonable to obtain some initial information, e.g. of a demographic nature, before presenting the actual BWS questions to the participants, the survey is structured by means of so-called question groups. In total, there are eleven question groups, each of which contains at least one, but in most cases several questions. Fig. 30 gives an overview of these question groups.

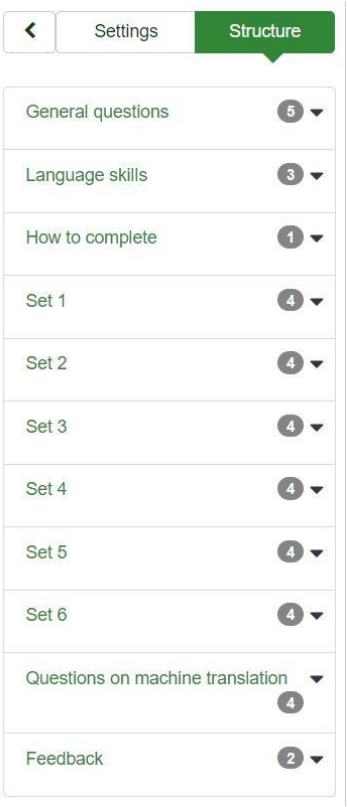


Figure 30: Question groups of the survey

As Fig. 30 shows, the first question group includes five general questions, two of which serve the purpose of ensuring that the participant belongs to one of the two intended target groups. The first of these two questions determines the participant's degree of expertise by asking for the highest educational degree they have completed. The second question asks for the field of study in which the participant obtained said degree, thus ensuring that the participant is either a translator or a natural scientist. The second question group enquires about the participant's first language as well as their level of skill in English and/or German. Since it can be assumed that one has a high language proficiency in one's own first language, it would be obsolete to ask participants to assess their language proficiency in their first language. Thus, the survey is designed to only ask for the English proficiency if the participant selects German

as their first language, and to only ask for the German proficiency if the participant selects English as their first language. Should a participant's first language be neither English nor German, they can select 'Other' and specify their first language in a comment field. In this case, however, the participant must assess their language proficiency both for English and for German.

The next question group does not actually contain a question, but rather a slide explaining how to correctly complete the subsequent question groups, namely Sets 1-6, which contain the BWS questions. The second to last question group includes questions about the participant's previous experience with and opinion of machine translation tools and their output. Finally, the last question group is an opportunity for participants to give feedback on the survey. The survey in its entirety is depicted in Appendix 3 of this thesis.

Using the example of Set 1, Fig. 31 shows how the BWS question groups have been set up in LimeSurvey.

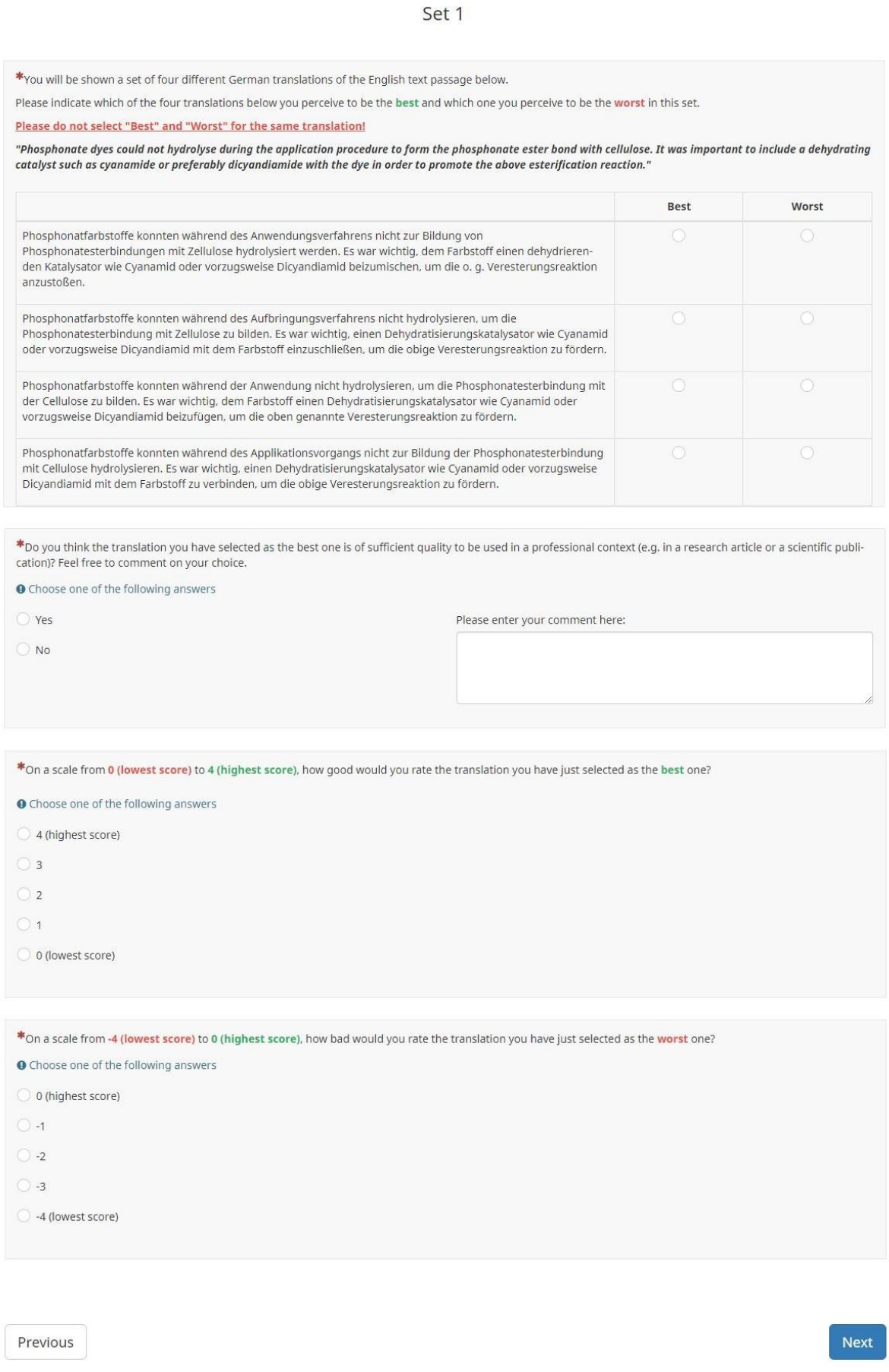


Figure 31: Design of BWS question groups in the survey

As can be seen in Fig. 31, the first question within each set is the actual best-worst scaling, where the participant compares the four translations to each other and makes their choice for the best and worst option. The question after that is intended to determine whether the participant thinks that the best option is good enough to be used in a professional context as opposed to the use for mere gisting purposes. Finally, in the last two questions, the participant must award plus points to the best option depending on how good they perceive it to be and minus points to the worst option depending on how bad they perceive it to be.

After the survey had been created, there was still one final step to be carried out before the survey could be distributed to potential participants, namely testing it. This test run was performed by an impartial person who had not seen the survey prior to testing it and who had never participated in a best-worst scaling before. That way, it could be ensured that the task at hand was clear and that the instructions were unambiguous and comprehensible even for someone who has no prior experience with the best-worst scaling method. The test run also served the purpose of checking whether there were any technical faults, logical errors, spelling/typing mistakes or other kinds of flaws in the survey that still needed to be corrected. Moreover, the test person was able to measure the time they needed to complete the survey without haste, which was then used as a time estimate in the welcome text of the survey. It is noted, however, that the participants did not have a time limit while completing the survey. After this test run, the survey was adapted corresponding to the feedback of the test person and was ready to be activated, i.e., to become accessible to participants.

5. Results

This chapter presents the data and results obtained from the survey that was conducted as part of this master's thesis. The data will be presented for each of the two target groups, and comparisons will be drawn between them. First, the participant profiles of both target groups will be discussed, then the results of the actual survey will be presented, and finally, the intra-rater reliability will be addressed.

5.1. Participant profiles

In total, there were nine translators and eight natural scientists who completed the survey, i.e., who answered every mandatory question and did not abandon the survey before they reached the end. Table 6 shows the profiles of the translators' group's participants while Table 7 shows the profiles of the natural scientists who took part in the survey.

Table 6: Participant profiles in the target group of the translators

	Age group	**Gender**	**Degree**	**Field**	**Exp. chem.**	**First language**	**EN prof.**
PT1	25-34	Female	Master's	Translation	1	German	C2
PT2	25-34	Male	Bachelor's	Translation	0	German	C1
PT3	25-34	Female	Bachelor's	Translation	0	German	C1
PT4	25-34	Female	Bachelor's	Translation	2	German	C2
PT5	35-44	Female	Diploma	Translation	2	German	C2
PT6	45-54	Female	Doctorate	Translation	1	German	C2
PT7	25-34	Female	Master's	Translation	0	German	C2
PT8	35-44	Female	Diploma	Translation	2	German	C2
PT9	25-34	Male	Master's	Translation	1	German	C2

Table 7: Participant profiles in the target group of the natural scientists

	Age group	Gender	Degree	Field	Exp. chem.	First language	EN prof.
PNS1	35-44	Female	Diploma	Chemistry	-	German	B2
PNS2	35-44	Male	Doctorate	Chemistry	-	German	C1
PNS3	35-44	Male	Doctorate	Chemistry	-	German	C2
PNS4	25-34	Female	Diploma	Pharmacy	2	German	C1
PNS5	25-34	Male	Bachelor's	Chemistry	-	German	C2
PNS6	25-34	Male	Master's	Physics	1	German	C2
PNS7	25-34	Female	Master's	Chemistry	-	German	C1
PNS8	35-44	Male	Diploma	Chemistry	-	German	C1

Table 6 and Table 7 show the participants' answers to the first question group ('General questions') and second question group ('Language skills') of the survey. The very first column lists the participants, where PT = 'Participant of Translation' and PNS = 'Participant of Natural Science'. The second column specifies which age group the respective participant belongs to, and the third column displays which gender the participant identifies as. The fourth column shows the highest educational degree of the participant, and the fifth column shows the academic field in which the participant obtained that degree. The sixth column specifies the participants' level of experience with specialised language in the field of chemistry, where 0 is the lowest and 4 is the highest level. The seventh column displays the respondents' first language, while the eighth and last column shows the participants' assessment of their own proficiency in the English language according to the Common European Framework of Reference for Languages (CEFR; see Council of Europe 2022a).

In the first group, 6 out of 9 PT, i.e., roughly 67%, belong to the age group 25-34. 2 out of 9 PT, which amounts to about 22%, are between 35 and 44 years old, and 1 out of the 9 PT, that is 11%, is between 45 and 54 years old. In the second group, exactly 50% of respondents, i.e., 4 out of 8 PNS, belong to the age group 25-34, while the other half is between 35 and 44 years old. Considering both groups, the most represented age group is thus 25-34 with more than half of all participants, namely 10 out of 17, or roughly 59%. The second largest age group is 35-44, with 6 out of 17 participants, or roughly 35%. 1 out of 17 participants, that is about 6%, belongs to the age group of 45-54. As far as gender is concerned, the majority of the participants in the translators' group, namely 7 out of 9 PT, or around 78%, are female, while only 2 out of 9 PT, or 22%, are male. In the group of the natural scientists, on the other hand, most participants, namely 5 out of 8 PNS, or around 63%, are male, while the remaining 3 PNS,

or 37%, are female. In total, i.e., across both groups, 10 out of 17 participants, namely 59%, are female and the remaining 7 participants, or 41%, are male.

Concerning the participants' educational degrees, it can be said that all 17 participants fulfil the requirement of having at least a bachelor's degree in their respective fields. Most of the participants even have a higher degree. In the translators' group, 3 out of 9 PT (roughly 33%) have a bachelor's degree, another 3 out of 9 PT (roughly 33%) have a master's degree, 2 out of 9 PT (roughly 22%) have a diploma, and 1 out of 9 PT (roughly 11%) has a doctorate. In the natural scientists' group, there is 1 out of 8 PNS (12.5%) with a bachelor's degree, there are 2 out of 8 PNS (25%) with a master's degree, 3 out of 8 PNS (37.5%) with a diploma, and 2 out of 8 PNS (25%) with a doctorate. Taking both groups into account, 4 out of 17 participants (roughly 24%) have a bachelor's degree, 5 out of 17 (roughly 29%) have a master's degree, another 5 out of 17 (roughly 29%) have a diploma, and 3 out of 17 (roughly 18%) have a doctorate. There are four different fields in which the participants obtained these degrees. In the group of the translators, of course, the field is always translation studies. In the natural science group, however, three different fields are represented. There are six participants from the field of chemistry, one from the field of pharmacy, and one from the field of physics.

Regarding the participants' experience with specialised language in the field of chemistry, 3 out of 9 PT (roughly 33%) indicated that they have no experience, another 3 out of 9 PT (roughly 33%) rated their experience a 1 on a scale from 0 (lowest) to 4 (highest), and the remaining 3 out of 9 PT (roughly 33%) specified that they had a medium level of experience, rated at 2. Since it is safe to assume that respondents who have a degree in the field of chemistry will automatically have a very high level of experience with the specialised language used in that domain, the survey was designed to only ask this question if the participant had chosen any field other than chemistry, which is why this question was disregarded for all PNS except for PNS4 and PNS6, who rated their experience levels at 2 and 1, respectively.

Concerning the language skills, all 17 participants happened to have German as their first language, which is why they were only asked to rate their English proficiency, not their German proficiency. In the translators' group, the vast majority, namely 7 out of 9 PT (roughly 78%), estimated their English skills to be at C2, the highest level of proficiency according to the CEFR. The remaining 2 PT (around 22%) indicated C1, which is the second highest level of proficiency. In the natural scientists' group, 3 out of 8 PNS (37.5%) rated their English skills at

C2, 4 out of 8 PNS (50%) at C1, and 1 out of 8 PNS (12.5%) at B2. Since both C1 and C2 users are categorised as “proficient users” (Council of Europe 2022b), it can be said that almost all participants (94%), with the exception of PNS1, are proficient users of English.

5.2. Results of the survey

As far as the data and results of the actual survey are concerned, Fig. 32 shows the results for the target group of the translators and Fig. 33 shows the results for the group of the natural scientists. Both figures exclude the responses obtained from Set 6, since these responses do not factor into the evaluation (as explained in Chapter 4.3.). They will be discussed separately in Chapter 5.3. within the framework of the intra-rater reliability.

Target text	1.1	1.2	1.3	1.4	PC	2.1	2.2	2.3	2.4	PC	3.1	3.2	3.3	3.4	PC	4.1	4.2	4.3	4.4	PC	5.1	5.2	5.3	5.4	PC
Tool/mode	H	G	D	S		M	H	G	D		S	M	H	G		D	S	M	H		G	D	S	M	
PT1	1	-3			Y	-1			3	Y			1	-3	N	3		-2		N		3	-2		Y
PT2	3		-3		Y	-2		3		Y		-3		4	Y	4		-2		Y		4		-3	Y
PT3		-2		3	N	-3			4	Y		-4		4	Y	1	-1			N		4		-4	Y
PT4	-3		2		N		3		-4	N			2	-2	N	1			-1	N		3		-2	N
PT5	4	-3			Y	0			4	Y		-3	3		Y		-1		3	Y		4		-3	Y
PT6		-3		3	N	-3		3		N		-3		3	N	3	-3			N			3	-3	N
PT7	3	-3			Y		-4		4	Y			-4	4	Y	4			-4	Y		4	-4		Y
PT8	4	-2			Y	-1	4			Y		-3	4		Y	3	-3			Y		3		-3	Y
PT9	4	-3			Y	-1	3			Y		-2	4		Y	4	-3			Y		4		-3	Y
Points	16	-19	-1	6		-11	6	6	11		0	-18	10	10		23	-11	-4	-2		0	29	-3	-21	

Figure 32: Results for the target group of the translators

Target text	1.1	1.2	1.3	1.4	PC	2.1	2.2	2.3	2.4	PC	3.1	3.2	3.3	3.4	PC	4.1	4.2	4.3	4.4	PC	5.1	5.2	5.3	5.4	PC
Tool/mode	H	G	D	S		M	H	G	D		S	M	H	G		D	S	M	H		G	D	S	M	
PNS1			2	-3	N	-2		3		Y	-2			4	Y			3	0	Y	3		-1		Y
PNS2	3	-2			Y	-1			4	Y	-2		4		Y		-3		3	Y		4		-4	Y
PNS3	-1		3		N	0			4	Y		-3		1	Y	3			-3	Y	-3	4			Y
PNS4	3	-3			Y	-1			2	Y		-2	3		Y		-2		3	Y		4		-4	Y
PNS5	4	-2			Y	-1	4			Y			0	3	Y			4	0	Y		4		-2	Y
PNS6	-2		3		Y	-3			4	Y			2	-4	Y	2			-2	Y	2	-3			Y
PNS7	3	-2			Y	0			3	Y	-3		3		Y	3	-2			Y		3		-4	Y
PNS8	3	-3			Y	-1			3	Y	-3		4		Y	3	-3			Y		4		-3	Y
Points	13	-12	8	-3		-9	4	3	20		-10	-5	16	4		11	-10	7	1		2	20	-1	-17	

Figure 33: Results for the target group of the natural scientists

The topmost line in Fig. 32 and Fig. 33 contains the target text numbers that have already been introduced in Table 5 (see Chapter 4.3.). Each group of four target texts represents a different set, i.e., all texts starting with the number 1 belong to Set 1, all texts starting with the

number 2 belong to Set 2, and so on. The second line in the two figures shows the translation mode or tool with which the respective target text was produced, where D = 'DeepL Translator', G = 'Google Translate', H = 'human', M = 'Microsoft Bing Translator', and S = 'SYSTRAN Translate'. Below the 'Tool/mode' cell, the very left column lists the participants. The participants' choices for the best option of each set are highlighted in green while their choices for the worst option of each set are highlighted in red. Moreover, the plus and minus points which the participants awarded to the options they selected are included in these figures as well. The column on the very right side of each individual set indicates whether the participant thought that the translation they selected as the best one would be good enough to be used in a Professional Context (PC), where Y = 'Yes' and N = 'No'. Finally, the bottommost line in the figures adds up the plus and minus points that were awarded to each target text in each set.

There are four sets in the survey in which there was a particularly high consensus among the participants concerning the best or worst choice. In Sets 4 and 5 in the translators' group and in Sets 2 and 5 in the natural scientists' group, DeepL Translator was rated particularly high, even considerably higher than the human translation. Both in Set 4 and Set 5 in the translators' group, 8 out of 9 PT, i.e., 89% of the group, perceived the translation produced by DeepL Translator as the best one. An even higher consensus was reached in Set 2 in the natural scientists' group, where all of the 8 participants, i.e., 100% of the group, agreed that the translation produced by Microsoft Bing Translator was the worst one. In an attempt to determine the reasons for the high agreement in these sets, Tables 8, 9 and 10 list the respective source texts and the translations that caused the high consensus among the respondents. Set 2 is compared in Table 8, Set 4 in Table 9, and Set 5 in Table 10.

Table 8: Reasons for high consensus in Set 2

Set 2	
Source text	Translation (by Microsoft Bing Translator)
When applied to nylon at pH 10, the dyes showed excellent substantivity, and yet at the boil the thiol-choline residue was eliminated to form the vinylsulphone dye which reacted readily with the deprotonated amines in nylon.	Wenn **sie** bei **pH 10** auf Nylon aufgetragen wurden, zeigten die Farbstoffe eine ausgezeichnete Substantivität, und doch wurde beim Kochen der Thiol-Cholin-Rückstand eliminiert, um den Vinylsulfonfarbstoff zu bilden, der leicht mit den deprotonierten Aminen in Nylon reagierte.

In Table 8, the parts of the respective translation which might have led to the consistently negative evaluations are highlighted in red. Altogether, there are not many parts in said translation that could be considered wrong, and yet the two instances at the very beginning of the sentence seem to have been sufficient for the respondents to deem it the worst translation out of the set. The translation begins with a cataphoric reference, meaning that the sentence is started with a pronoun (‘sie’) that refers to a subject which has not been introduced yet at that point (‘die Farbstoffe’). In non-literary texts and, more specifically, in highly scientific texts like these ones, where unambiguity and comprehensibility are of particular importance, literary and stylistic devices like cataphora can disrupt the reading flow and are therefore inappropriate in this context. Since none of the other translations of Set 2 feature cataphora, it can be assumed that this is the main reason for the many negative evaluations.

There is still one other part that is highlighted in red in Table 8, namely ‘pH 10’. While it is common in English to talk about the pH scale as merely ‘pH’ (as opposed to ‘pH value’, ‘pH number’ or similar), the matter is different in German, where this scale is usually referred to as ‘pH-Wert’ (‘pH value’). The German formulation ‘bei pH 10’ that was used in this translation by Microsoft Bing Translator thus sounds rather unnatural, which might have been an additional contributing factor for the high agreement that this translation was the worst one in the set.

Table 9: Reasons for high consensus in Set 4

Set 4	
Source text	Translation (by DeepL Translator)
With regard to heterocyclic compounds as coupling components, the importance of the formerly widespread pyrazolone-, aminopyrazole-, and 4-hydroxyquinolone-based yellow azo dyes has greatly diminished with the advent of the tinctorially superior and therefore more economical pyridone azo dyes.	Bei den **heterozyklischen** Verbindungen als Kupplungskomponenten hat die Bedeutung der **früher weit verbreiteten** gelben Azofarbstoffe **auf** Pyrazolon-, Aminopyrazol- und 4-Hydroxychinolon-**Basis** mit dem Aufkommen der **farblich überlegenen** und damit **kostengünstigeren** Pyridon-Azofarbstoffe stark abgenommen.

Table 9 compares the source text for Set 4 with the translation that was considered the best one by almost all participants. The parts of the translation that are particularly well translated, especially in contrast to the other translations, are highlighted in green. The first instance is the word ‘heterozyklisch’, which was spelled in a less common spelling variant in some of the other translations, namely as ‘heterocyclisch’. The next part that is marked as correct is ‘früher weit verbreiteten’, which is the correct translation for ‘formerly widespread’. In some of the other options in that set, this part was translated as ‘bisher weit verbreiteten’ (‘widespread up to now’), which does not convey quite the same meaning as the source text.

The next part that is better in this translation than in the other ones is the construction ‘auf […] -Basis’. Due to the use of nominalisation (‘-Basis’), which is used very often in the German language, the passage sounds more authentic and reads less like a translation than it would have if the participle (‘-basierten’) had been used instead. Finally, the two adjectives ‘farblich überlegenen’ (‘tinctorially superior’) and ‘kostengünstigeren’ (‘more economical’) sound more natural than some of the other translations that occurred in that set. Summing up, it can be assumed that the combination of all of these natural-sounding grammatical constructions and correct equivalents resulted in an overall adequate and satisfying translation, which was responsible for the high agreement among the participants that this translation was the best one in the set.

Table 10: Reasons for high consensus in Set 5

Set 5	
Source text	Translation (by DeepL Translator)
These biological sorbents were found to be efficient in binding with basic dyes rather than acid dyes. This was attributed mainly to the Coulombic attraction between the negative surface of the adsorbent and the positively charged ions of basic dyes.	**Es wurde festgestellt, dass** diese biologischen **Sorptionsmittel** eher basische Farbstoffe als saure Farbstoffe binden. Dies wurde hauptsächlich auf die **Coulombsche Anziehungskraft** zwischen der negativen Oberfläche des **Adsorptionsmittels** und den positiv geladenen Ionen der **basischen Farbstoffe** zurückgeführt.

In Table 10, the source text for Set 5 and the translation that led to the high consensus are contrasted. Again, the parts of this target text that were translated particularly well are highlighted in green. On the one hand, there are some instances that exhibit good terminological equivalence, for instance ‘Sorptionsmittel’ for ‘sorbents’, ‘Coulombsche Anziehungskraft’ for ‘Coulombic attraction’ and ‘Adsorptionsmittel’ for ‘adsorbent’. Moreover, ‘basic dyes’ was correctly translated as ‘basische Farbstoffe’ in this translation. In some of the other options in the set, this was translated as ‘Basisfarbstoffe’, which is not a common or correct equivalent. Furthermore, in the beginning of the first sentence of the translation, DeepL did not translate literally, but opted for a slight rephrasing that makes the translation sound more natural. The English construction ‘were found to be’ is quite difficult to translate literally into German, which is why ‘Es wurde festgestellt, dass’ (‘It was determined that’) sounds less ambiguous and therefore clearer, which is preferable in a highly scientific context like this.

Another aspect that is noticeable from the results is that some participants across both target groups did not always seem to have sound reasoning for their responses to the question whether they considered the translation they selected as the best one to be good enough to be used in a professional context. In some cases, these responses even seem to be rather contradictory. For instance, PT1 chose ‘Yes’ for the best translation in Set 1 although they only awarded it 1 plus point, which would mean that they did not actually think the translation was altogether good. The same applies to PNS3 in Set 3. Conversely, PT1 chose ‘No’ in Set 4 even though they gave 3 out of 4 possible plus points, i.e., the second-best score, to the translation

in question. Other instances where a participant indicated ‘No’ while simultaneously awarding the second-best point score to the translation include PT3 in Set 1, PT4 in Sets 2 and 5, PT6 in all sets, or PNS3 in Set 1. Another phenomenon that was observed quite often is that some participants either selected always ‘Yes’ or always ‘No’ throughout the entire survey. The participants who always selected ‘Yes’ were PT2, PT5, PT7, PT8, PT9, PNS2, PNS4, PNS5, PNS6, PNS7, and PNS8. The participants who always selected ‘No’ were PT4 and PT6.

There are a few possible reasons for these response patterns. One of the explanations for the cases in which the ‘Yes’/’No’ answers seemingly contradicted the point score could be that the participants in question might have focussed more on particular aspects that worked well in the translation, e.g. terminology or fluency, and might therefore have been more forgiving in terms of any grammatical or syntactical errors. Another reason could be that the question about the professional context simply carried less weight in their decision-making process than the point score did. Regarding the participants who continuously selected ‘No’ across all sets, a possible explanation could be that since the participants knew that NMT tools were involved, they might have selected ‘No’ per default because they might hold the opinion that no NMT tool could ever produce a translation of sufficient quality for a professional context. A reason why so many participants continuously selected ‘Yes’ could, once again, be that the question about the professional context carried less weight in the participants’ decision-making process.

Returning to the scores of the BWS, Table 11 and Table 12 show the total points that were awarded to each translation mode or tool by each PT and PNS, respectively.

Table 11: Total points awarded by the translators

	D	G	H	M	S
PT1	9	-6	2	-3	-2
PT2	5	7	3	-10	0
PT3	9	2	0	-11	2
PT4	2	-2	1	-2	0
PT5	8	-3	10	-6	-1
PT6	3	3	0	-9	3
PT7	12	1	-9	0	-4
PT8	6	-2	12	-7	-3
PT9	8	-3	11	-6	-3
Total	62	-3	30	-54	-8

Table 12: Total points awarded by the natural scientists

	D	G	H	M	S
PNS1	2	10	0	1	-6
PNS2	8	-2	10	-5	-5
PNS3	14	-2	-4	-3	0
PNS4	6	-3	9	-7	-2
PNS5	4	1	8	1	0
PNS6	6	-2	-2	-3	0
PNS7	9	-2	6	-4	-5
PNS8	10	-3	7	-4	-6
Total	59	-3	34	-24	-24

Using the same abbreviations for the translation tools/mode as well as for the participants that were already introduced earlier, Table 11 and Table 12 show how the individual objects were rated in total by each participant across all five sets. Positive points are highlighted in green, negative points are highlighted in red, and neutral evaluations are coloured grey. What is most noticeable from these tables is that DeepL Translator is the only object that received a positive score by each and every single participant. This does not mean that DeepL Translator was never selected as the worst option. It merely means that, across all five sets, DeepL Translator achieved more positive points than negative points. Moreover, these two tables show that DeepL Translator and the human translation are the only two objects that attained positive total values. Remarkably, DeepL Translator obtained significantly more points than the human translation in both target groups.

In addition to the total points, there are also other scores that can be obtained, such as the average points or how often the objects were selected as the best and worst option. Table 13 shows these scores for each object, i.e., for each translation mode/tool, in the target group of the translators, while Table 14 shows the same types of scores for the target group of the natural scientists.

Table 13: Scores for the five objects in the target group of the translators

Tool/mode	Total points	Total points/ no. of part.	Best	Worst	B-W
DeepL Translator (D)	62	6.89	21	2	19
Human (H)	30	3.33	15	5	10
Google Translate (G)	-3	-0.33	6	9	-3
SYSTRAN Translate (S)	-8	-0.89	3	7	-4
Microsoft Bing Translator (M)	-54	-6	0	22	-22

Table 14: Scores for the five objects in the target group of the natural scientists

Tool/mode	Total points	Total points/ no. of part.	Best	Worst	B-W
DeepL Translator (D)	59	7.38	19	1	18
Human (H)	34	4.25	13	7	6
Google Translate (G)	-3	-0.38	6	7	-1
SYSTRAN Translate (S)	-24	-3	0	10	-10
Microsoft Bing Translator (M)	-24	-3	2	15	-13

The first column of Tables 13 and 14 lists the objects, i.e., the translation tool or mode of translation, represented in the same colours they were assigned in Table 4 (see Chapter 4.3.). The second column adds up the total points, i.e., the sum of all plus and minus points which were awarded to the respective object during the survey, while the third column displays the average points that were awarded to it, i.e., the value that results from dividing the object's total points by the number of participants in the target group. The fourth and fifth columns represent the number of times the object in question was selected as the best option of the set and as the worst option of the set, respectively. Finally, the sixth and last column is the score that is obtained by subtracting the number of times the object was picked as the worst from the number of times it was picked as the best.

What is noteworthy about Tables 13 and 14 is that the order of the objects' ratings is the same in both target groups. It can be gathered from these two tables that DeepL Translator, remarkably, achieved the highest scores in both groups, not only in terms of the awarded total points, namely 62 in the translators' group and 59 in the natural scientists' group, but also regarding the B-W score, which is 19 in the translators' group and 18 in the natural scientists' group. With total points of 30 and 34 and B-W scores of 10 and 6, the human translation occupies the second place in both groups. The third place goes to Google Translate, which achieved total points of -3 in both groups and B-W scores of -3 and -1. SYSTRAN Translate reached the fourth place with -8 and -24 points and B-W scores of -4 and -10. Finally, Microsoft

Bing Translator received the lowest scores, namely -54 points in the translators' group and -24 points in the natural scientists' group, as well as B-W scores of -22 and -13. If only the awarded plus and minus points were considered, Microsoft Bing Translator would share the last place with SYSTRAN Translate in the group of the natural scientists, since both have -24 points. But seeing as Microsoft Bing Translator received a lower B-W score than SYSTRAN Translate, Microsoft Bing Translator is ranked lower than SYSTRAN Translate.

Finally, the results of the question group 'Questions on machine translation', which deals with the participants' previous experiences with MT tools, are presented in Table 15 for the translators' group and in Table 16 for the natural scientists' group.

Table 15: Results of the MT questions in the target group of the translators

	Used MT tools before?	**Frequency**	**Satisfaction**
PT1	Yes	Often	Somewhat satisfied
PT2	Yes	Sometimes	Somewhat satisfied
PT3	Yes	Sometimes	Somewhat satisfied
PT4	Yes	Sometimes	Somewhat dissatisfied
PT5	Yes	Sometimes	Somewhat satisfied
PT6	Yes	Sometimes	Somewhat dissatisfied
PT7	Yes	Sometimes	Somewhat satisfied
PT8	Yes	Sometimes	Neutral
PT9	Yes	Rarely	Neutral

Table 16: Results of the MT questions in the target group of the natural scientists

	Used MT tools before?	**Frequency**	**Satisfaction**
PNS1	Yes	Sometimes	Neutral
PNS2	Yes	Rarely	Neutral
PNS3	Yes	Rarely	Somewhat satisfied
PNS4	Yes	Rarely	Neutral
PNS5	No	-	-
PNS6	Yes	Rarely	Somewhat satisfied
PNS7	Yes	Sometimes	Somewhat satisfied
PNS8	Yes	Sometimes	Somewhat satisfied

On the very left of Tables 15 and 16, the first column lists the participants with their respective abbreviations. The second column displays the participants' indication about whether or not they had already used MT tools before. The third column provides the participants' answers to the question how often they use MT tools on a scale from 'Never' to

‘Very often’. Finally, the fourth and last column reflects how satisfied the participants usually are with the output of MT tools. The available answers for this question were ‘Very dissatisfied’, ‘Somewhat dissatisfied’, ‘Neutral’, ‘Somewhat satisfied’ and ‘Very satisfied’.

As can be gathered from these two tables, all 9 PT, i.e., 100% of the translators, and 7 out of 8 PNS, i.e., roughly 88% of the natural scientists, had already used an MT tool before. In total, 16 out of 17 participants, which amounts to about 94%, were already familiar with MT tools. 7 out of 9 PT, or 78%, indicated that they sometimes used MT tools, 1 out of 9 PT, or 11%, indicated that they often used MT tools, and another 1 out of 9 PT, or 11%, indicated that they rarely used them. In the second target group, 4 out of 8 PNS, i.e., exactly 50%, indicated that they rarely used MT tools, while 3 out of 8 PNS, or around 38%, indicated that they sometimes used them. Thus, the majority of all participants, namely 10 out of 17, or around 59%, sometimes uses MT tools, while 5 out of 17, or 29%, rarely use them and only 1 out of 17, or 6%, often uses them. In terms of the participants’ satisfaction with the output of MT tools, 5 out of 9 PT, or 56%, and 4 out of 8 PNS, or 50%, are somewhat satisfied with MT output. Across both groups, this means that more than half of all participants, namely 9 out of 17, around 53%, are somewhat satisfied. Only 2 out of 9 PT, that is 22% of the group or 12% of all participants, deem the output of MT tools as somewhat dissatisfying. The remaining participants, namely 2 out of 9 PT (22%) and 3 out of 8 PNS (38%), amounting to roughly 29% of all participants, have neutral feelings about the output of MT tools.

Not only do the translators tend to use MT tools more often than the natural scientists, but they also seem to be more critical when assessing the output of MT tools. This is most likely due to the fact that professional translators, having immersed themselves in translation studies for years, have a different approach to evaluating translations than non-translators. Trained translators might pay attention to aspects of a translation that might go unnoticed by persons who have never dealt with translation studies.

5.3. Intra-rater reliability

As was briefly explained in Chapter 4.3., the intra-rater reliability was tested in this survey by having the participants evaluate one of the sets twice. For this purpose, Set 2 was repeated after Set 5, where it was renamed to Set 6. The results of this reliability test, i.e., the comparison

between the responses obtained from Set 2 and Set 6, are shown in Table 17 for the translators' group and in Table 18 for the natural scientists' group.

Table 17: Intra-rater reliability in the target group of the translators

Target text	2.1	2.2	2.3	2.4	PC	6.1	6.2	6.3	6.4	PC
Tool/mode	M	H	G	D		M	H	G	D	
PT1	-1			3	Y	-1			3	Y
PT2	-2		3		Y	-3		3		Y
PT3	-3			4	Y	-3	4			Y
PT4		3		-4	N		-2	2		N
PT5	0			4	Y	0			3	Y
PT6	-3		3		N	-3		3		N
PT7		-4		4	Y	-3			4	Y
PT8	-1	4			Y	-1	4			Y
PT9	-1	3			Y	-1	3			Y
Points	-11	6	6	11		-15	9	8	10	

Table 18: Intra-rater reliability in the target group of the natural scientists

Target text	2.1	2.2	2.3	2.4	PC	6.1	6.2	6.3	6.4	PC
Tool/mode	M	H	G	D		M	H	G	D	
PNS1	-2		3		Y	-2		3		N
PNS2	-1			4	Y	0			4	Y
PNS3	0			4	Y		0	2		Y
PNS4	-1			2	Y	-1			3	Y
PNS5	-1	4			Y		4		-1	Y
PNS6	-3			4	Y	-2			3	Y
PNS7	0			3	Y	0			4	Y
PNS8	-1			3	Y	0			3	Y
Points	-9	4	3	20		-5	4	5	16	

Table 17 and Table 18 compare the answers given in Set 2 and Set 6 in the respective target groups. Once again, the abbreviations and the numbering which are already known from previous chapters are used for the target texts, translation tools/mode, and participants. The rows that remain white indicate the participants who fully passed the validation test, meaning that they selected the exact same options as best/worst, awarded the exact same points, and gave the exact same response to the question about the professional context in both sets. The light red rows indicate the participants who passed the validation test regarding their selection for the best/worst option, but who either awarded a slightly different point score to the translations

or changed their mind concerning the question about the professional context. Last but not least, the rows that are highlighted in dark red indicate the participants who failed the validation test, i.e., whose best/worst selections from Set 6 do not correspond to their selections from Set 2. Some of the participants who failed did not only fail in terms of the best/worst option, but also with regard to the awarded number of points between the two sets. The bottom row of both figures sums up the points per object in each of the two sets.

In the translators' group, 3 out of 9 PT (roughly 33%) failed the test, resulting in a self-consistency of roughly 67%. In the group of the natural scientists, 2 out of 8 PNS, or 25%, failed the test, which leads to a self-consistency of 75%. Thus, it can be said that the best-worst scaling proved to be a reliable research method in about two thirds to three quarters of the cases in this thesis. In the translators' group, the order in which the objects were ranked would have stayed the same, had the results from Set 6 been taken into account instead of the responses obtained from Set 2. In the natural scientists' group, it would have made a small impact, namely that Microsoft Bing Translator would have had four points more, which would have put it in the fourth place instead of the fifth.

6. Discussion

This chapter discusses the results obtained from the survey as well as their significance and implications in various regards. First, Chapter 6.1. will make reference to the research questions and assumptions for this thesis that were outlined in Chapter 1.1. It will provide a tendency regarding the research questions and examine to what extent the actual results correspond to or deviate from the assumptions that have been made. Next, Chapter 6.2. aims to explain the results and their implications in more detail, while Chapter 6.3. will discuss the correlation of this study's results to the related work that was introduced in Chapter 3. Finally, Chapter 6.4. will present the feedback that was gathered from the participants and elaborate on the author's own experiences and insights gained from the survey. It will also discuss the benefits and shortcomings of the method that were determined in the course of the empirical research process.

6.1. Revisiting the research questions and assumptions

With regard to the first research question, it can be said that, although both the translators and the natural scientists rated the raw output of three out of the four examined NMT tools worse than the human translation, there is one NMT tool that managed to perform significantly better than human translation, namely DeepL Translator. It is also the only NMT tool which achieved a positive total score, while the other three tools all received a negative total score in both target groups. The assumption for the first research question was that the participants of the evaluation would, on average, pick the human translation as the best one more often than the machine translation and that the human translation would be perceived as exhibiting higher quality than the machine translation. Astonishingly, this assumption did not prove to be true, seeing as the translations by DeepL Translator were picked as the best option more often than the human translations and DeepL Translator also received more total points than the human translation mode by each participant. This leads to the conclusion that the translations produced by DeepL Translator were perceived as exhibiting an even higher quality than the human translations.

For the second research question, the assumption made in the beginning was that among the online neural machine translation tools that were examined in the survey (namely DeepL Translator, Google Translate, Microsoft Bing Translator and SYSTRAN Translate), DeepL Translator would generate the most useful translations, i.e., translations that would be rated

better than those produced by the other NMT tools. This assumption undoubtedly proved to be true, seeing as DeepL Translator scored the most points by far, not only out of the four NMT tools, but out of all five objects, thus even outperforming human translation.

6.2. Explanation of the results

There are many possible explanations for the fact that DeepL Translator performed so well in this survey. Firstly, DeepL Translator has been known to create better translations than most of its competitors due to what is described by the company itself as a different topology in the networks that DeepL Translator uses (DeepL 2021), i.e., differences in the number of layers and nodes within the network and differences in the arrangement of these elements. Secondly, the company has focussed on acquiring special training data, which, in turn, leads to a higher quality of the resulting translations (DeepL 2021). This special training data could possibly include a higher percentage of research papers on natural science, which might also be the reason why DeepL Translator seems to be particularly adequate and well-suited for the text domain of natural science that was examined in this survey. Another reason for their success, according to DeepL (2021), is the fact that the company not only makes use of the training methods that are conventional for MT, but also applies techniques that have been borrowed from other machine learning areas.

After examining what sets DeepL Translator apart from the other NMT tools, there are still other reasons that might influence the performances of the tools in this particular research. For one, it can be assumed that the high quality of the translations is directly linked to the languages that were examined, namely English and German, which is a rather common language combination. As explained earlier in this thesis, English is the most common and most dominant language online, with more than a quarter of all people on the internet speaking or using it as of January 2020 (Statista 2020). Therefore, a considerably large portion of the training data that is used for training NMT tools is in English. And although the German language is encountered far less often on the internet than English, the two languages are closely related to each other from a linguistic perspective, seeing as they are both Germanic languages and share similar structures and grammatical patterns. This consequently leads to a better MT output. In other words, machine translation between English and German is bound to yield far

better results than machine translation between two languages that are, linguistically speaking, very different from each other or that are not related at all.

As determined in the results, DeepL Translator performed well with regard to terminology, adequacy and fluency, and it produces output that sounds natural in the target language. All in all, participants largely indicated that they perceived the translations by DeepL Translator to be good enough to be used in a professional context, i.e., for purposes of publication, and there were only few cases in which DeepL Translator's translations were deemed to be insufficient for use in a professional context. In summary, the most important lessons learned from the results of the research conducted in this thesis are the following. Firstly, DeepL Translator achieved the best results out of all five options in this survey and thus, surprisingly, even surpassed human translation. This fact points towards the conclusion that NMT is indeed able to outperform HT without human intervention under certain conditions. However, this does not mean that all NMT tools can be considered equal to or better than human translation, and it does not necessarily mean that NMT can outperform HT in any domain. Secondly, the remaining three NMT tools, namely Google Translate, Microsoft Bing Translator, and SYSTRAN Translate, all achieved a negative point score as well as a negative B-W score, meaning that they all were selected as the worst option more often than they were selected as the best one. This leads to the conclusion that most NMT tools still require post-editing in order to produce adequate and usable translations.

6.3. Correlation to related work

This chapter links the results of the survey carried out as part of this thesis to the results of the related work that was examined in Chapter 3. Although the objectives of the research presented in Chapter 3 are very similar to the objectives of this thesis in that they all compare MT against HT, none of the studies from Chapter 3 conducted their research with the exact same research design as this thesis.

For instance, while this thesis used the best-worst scaling, most of the related research was conducted using other human as well as automatic research and evaluation methods. Ahrenberg (2017) analysed the number of shifted/unshifted clauses and used the BLEU and TER scores, Läubli et al. (2018) applied pairwise ranking, Popel et al. (2020) opted to use human error annotation, and Munková et al. (2019) used their own tool that featured different

metrics for automatic MT evaluation. The two languages examined within the framework of this thesis, namely English and German, are each involved in some of the related work, but never together. English, for instance, is involved in Ahrenberg (2017), Läubli et al. (2018), Popel et al. (2020), and Vasheghani Farahani (2020), while German is involved in Munková et al. (2019).

As far as the results are concerned, it can be said that the results of this thesis partially correspond to and partially contradict the results from the related work. Ahrenberg found that MT did not yet yield results that were good enough for publication standards and that humans produced stylistically better sentences than MT (2017: 25). The results of this thesis contradict Ahrenberg's findings, since the output of DeepL Translator was largely considered of sufficient quality to be used in publications. As far as style is concerned, DeepL Translator was able to match HT, but the other NMT tools mostly were not. Munková et al. (2019: 231) determined in their study that, as far as quality is concerned, HT achieved better results than MT, but similar results to the process of post-editing MT. Again, the results presented in this thesis do not confirm the first part of these findings, since one NMT tool did in fact achieve better results than HT, but no comments can be made with regard to the second part, since post-editing was not examined as part of this thesis. The results from Läubli et al.'s study showed that HT and MT did not show significant differences in isolated sentences, but at the document level, HT was perceived as noticeably better than MT (2018: 4793). These findings cannot be confirmed by the results of this study, since a) this thesis did not take whole documents into account, and b) HT and MT did in fact show significant differences in quality in isolated sentences, as can be seen in the results. Popel et al. (2020) found in their study that NMT was able to approach the quality of HT and even performed better than HT with regard to spelling, grammar, and, above all, with regard to adequacy, but that HT, on the other hand, made fewer fluency errors and context-related errors. Although the study conducted in this thesis did not apply error categories, it can be said that one of the NMT tools examined in this thesis even outperformed HT, which is why the findings from Popel et al.'s study (2020) can be confirmed. Finally, Vasheghani Farahani determined that MT was indeed capable of producing translations whose adequacy could be equal to HT, but that the end product mostly still required human post-editing to make up for other errors made by MT (2020: 101). Again, these findings can be confirmed by the research conducted in this thesis because one MT tool was able to produce translations that matched and even outperformed HT without the need for post-editing.

6.4. Feedback and insights

As was already mentioned at an earlier point, at the time of writing this thesis, the best-worst scaling method has not been used very often in research conducted in the field of translation studies. This subchapter will present the feedback from the participants that was obtained at the end of the survey, and it will discuss the author's own experiences and insights concerning the method and the survey in general.

As far as positive feedback is concerned, some participants noted that they thought the survey was well-prepared and structured in a clear and unambiguous manner. Some participants explicitly noted that they thought this method of asking questions, i.e., the best-worst scaling, was unconventional and unusual, yet interesting and clear. Other participants made positive remarks about the subject of the survey in general, commenting that they considered the survey's research objective to be topical and interesting.

On the other hand, there were also some responses containing negative feedback. About half of all participants, which amounts to almost all of the negative feedback that was given, commented that they found it tiring and tedious to have to read and compare so many passages in such detail and in one sitting, and that they lost focus during the course of the survey. Since losing focus is a common problem in human evaluation methods, this was an issue that was already anticipated to some extent by the author of this thesis. Apart from the problem of the large text volumes, one participant in the translators' group also remarked that it was difficult to judge the translations without the corresponding expertise in the field of chemistry.

Although the received negative feedback repeatedly deals with the problem of high text volumes and participants losing focus, the author of this thesis would decide against using shorter or fewer texts if she were to repeat the study. This is due to the fact that three out of the five source text passages that were used only consisted of one sentence as it is, and the remaining two source text passages also only featured two sentences. In any case, none of the source text passages consisted of more than 45 words, which is why it would be difficult to choose even shorter passages without sacrificing the high degree of specialised language. Moreover, it would not be advisable in terms of the BWS method to include fewer source text passages, since five is already on the lower end of the scale. The only possibility to reduce the

text volume that the author would consider feasible would be to eliminate the repeated question that was used to test the intra-rater reliability.

In general, the chosen method could be deemed appropriate for this kind of experiment. It proved to be simple to use and evaluate, and it produced clear, unambiguous results that were easy to interpret. However, the author found that one drawback of using BWS is that the use of BIBDs, while undoubtedly convenient in cases that feature a large number of objects, can be somewhat restrictive in studies that only examine a small number of total objects, since this leads to significantly fewer and less varied options for how a BIBD can be constructed. Regardless, the author considers BWS to be an interesting and effective method for measuring subjective estimations and perceptions such as the quality of different translations.

7. Conclusion

This master's thesis was written with the objective of determining how translators as well as natural scientists rate the raw output of neural machine translation in comparison to human translation in highly specialised, domain-specific texts taken from the field of chemistry and in the language direction English to German without being explicitly informed about the type of translation. Furthermore, this thesis aimed to determine which of the commercially available online neural machine translation tools used in this experiment yields the best results in terms of a numeric rating without any human intervention such as pre- or post-editing.

This was achieved by extracting five English-language source text passages from highly specialised and domain-specific literature from the field of chemistry and producing one human translation as well as four machine translations of each of these passages. The machine translations were produced using four different online neural machine translation tools, namely DeepL Translator, Google Translate, Microsoft Bing Translator, and SYSTRAN Translate. By means of a best-worst scaling, in which participants were shown sets containing four translations each and were asked to select the best and the worst one, the translations were compared and evaluated against each other. Moreover, the participants were asked to award plus and minus points to each option they selected as the best/worst one, depending on how good/bad they thought it was. The result was a numeric score that was counted for each of the translation methods.

The findings of this survey were that both the translators and the natural scientists rated one of the NMT tools, namely DeepL Translator, significantly better than all of the other objects, including even the human translation. DeepL Translator received by far the best score among all examined translation tools/modes, and it was also the only NMT tool that achieved a positive numeric rating. From that, it can be concluded that it is possible for machine translation tools to reach or even exceed the quality level of a human translation under certain conditions without the need for human post-editing.

As far as future research in the area of machine translation quality is concerned, there are several aspects that are still worth exploring, or regards in which a study like this could be done differently. For instance, it would be interesting to conduct a study like this with languages that have different grammatical patterns, that are used in very different socio-cultural markets, or that use different alphabets. It would also be worthwhile to examine the quality of MT if the languages involved are not as well-represented online as English. Another interesting aspect

would be a different subject area, i.e., to see how MT performs when translating texts from other highly specialised fields, not only within the natural sciences, but also within other domains, such as health sciences, engineering, social sciences, agriculture, or humanities. Another direction for future research could be to incorporate more than one human translation in the BWS, i.e., to ensure that the same number of human translations and machine translations are compared and assessed. Finally, the results of studies on this topic would also benefit from a larger group of participants, as it would make the study more representative. If the study were conducted on a larger scale, it might be possible to spread the survey more widely to specialised communities, or to attract more participants by offering prizes or a monetary incentive as a reward for participating in and completing the survey.

The future of machine translation seems promising, that much is clear. Machine translation, particularly neural machine translation, is a technology that has undeniably gained in importance significantly and that has seen astounding progress both in the past decades and even in the past few years. As recent research as well as the results of this study have shown, the raw output of machine translation tools can reach a level of quality equal to or even better than that of human translation in certain situations and under certain conditions. Given the fact that AI and deep learning are fields that are evolving at an extremely fast pace, the topic of machine translation quality will continue to be relevant and studies in this regard might produce even more promising results in as little as a few years. Thus, it remains to be seen whether the decades-old claim that machine translation will be ready to replace human translators within a few years will ultimately come true.

Bibliography

Ahrenberg, Lars (2017). Comparing Machine Translation and Human Translation: A Case Study. In: Temnikova, Irina; Orasan, Constantin; Corpas, Gloria; Vogel, Stephan (eds). *RANLP 2017 The First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT) Proceedings of the Workshop.* Linköping University, Sweden, 21-28.

Al-Degs, Yahya; Khraisheh, Majeda A. M.; Allen, Stephen J.; Ahmad, Mohammad N. (2000). Effect of Carbon Surface Chemistry on the Removal of Reactive Dyes from Textile Effluent. *Water Research* 34 (3), 927-935.

ALPAC (1966). *Languages and Machines: Computers in Translation and Linguistics.* A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.

Banerjee, Satanjeev; Lavie, Alon (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor*, 65-72.

Bornmann, Lutz; Mutz, Rüdiger (2015). Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publications and Cited References. *Journal of the Association for Information Science and Technology* 66 (11), 2215-2222.

Burchardt, Aljoscha; Porsiel, Jörg (2017). Vorwort: Was kann maschinelle Übersetzung und was nicht? In: Porsiel, Jörg (ed.). *Maschinelle Übersetzung: Grundlagen für den professionellen Einsatz*. Berlin: BDÜ Fachverlag, 11-17.

Burchardt, Aljoscha; Macketanz, Vivien; Dehdari, Jon; Heigold, Georg; Peter, Jan-Thorsten; Williams, Philip (2017). A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. *The Prague Bulletin of Mathematical Linguistics* 108, 159-170.

Callison-Burch, Chris; Osborne, Miles; Koehn, Philipp (2006). Re-Evaluating the Role of BLEU in Machine Translation Research. In: *11th Conference of the European Chapter of the Association for Computational Linguistics*, 249-256.

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

Childs, Peter E.; Markic, Silvija; Ryan, Marie C. (2015). The Role of Language in the Teaching and Learning of Chemistry. In: García-Martínez, Javier; Serrano-Torregrosa, Elena (eds). *Chemistry Education: Best Practices, Opportunities and Trends.* Weinheim: Wiley-VCH, 421-445.

Council of Europe (2022a). Common European Framework of Reference for Languages (CEFR). https://www.coe.int/en/web/common-european-framework-reference-languages (Accessed: 30 March 2022)

Council of Europe (2022b). Global scale – Table 1 (CEFR 3.3): Common Reference levels. https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale (Accessed: 30 March 2022)

DeepL (2021). How Does DeepL Work? https://www.deepl.com/en/blog/how-does-deepl-work (Accessed: 13 April 2022)

DeepL Translator (2022). https://www.deepl.com/en/translator (Accessed: 20 March 2022)

Doherty, Stephen (2017). Issues in Human and Automatic Translation Quality Assessment. In: Kenny, Dorothy (ed.). *Human Issues in Translation Technology*. London: Routledge, 131-148.

European Commission (2003). *Third European Report on Science & Technology Indicators.* Luxembourg: Office for Official Publications of the European Communities. http://www.eurosfaire.prd.fr/7pc/doc/1124294203_third_european_report_on_science_technology_indicators_2003.pdf (Accessed: 2 July 2021)

Fiederer, Rebecca; O'Brien, Sharon (2009). Quality and Machine Translation: A Realistic Objective? *The Journal of Specialised Translation* 11, 52-74.

Finetti, Fiorenza (2020). *Pre-Editing als Vorschlag zur Verbesserung neuronaler maschineller Übersetzung.* Masterarbeit, Universität Wien.

Forcada, Mikel L. (2017). Making Sense of Neural Machine Translation. *Translation Spaces* 6 (2), 291-309.

Gaspari, Federico; Hutchins, William John (2007). Online and Free! Ten Years of Online Machine Translation: Origins, Developments, Current Use and Future Prospects. *Proceedings of MT Summit XI, Copenhagen, Denmark,* 199-206.

Google Translate (2022). https://translate.google.com/?hl=en (Accessed: 20 March 2022)

Göpferich, Susanne (2009). Comprehensibility Assessment Using the Karlsruhe Comprehensibility Concept. *The Journal of Specialised Translation* 11, 31-52.

Göschl, Janina (2022). *Die Auswirkung von Pre-Editing-Methoden auf die Translationsqualität von maschinell übersetzten Texten und deren Einschätzung durch Sprachexpert*innen im Vergleich zu Lai*innen*. Masterarbeit, Universität Wien.

Gouadec, Daniel (2010). Quality in Translation. In: Gambier, Yves; Van Doorslaer, Luc (eds). *Handbook of Translation Studies: Volume 1.* Amsterdam/Philadelphia: John Benjamins Publishing Company, 270-275.

Gwet, Kilem L. (2014). Intrarater Reliability. In: Balakrishnan, N. (ed.). *Methods and Applications of Statistics in Clinical Trials: Volume 2.* Hoboken, New Jersey: John Wiley & Sons, 340-356.

Hess, Stephane; Daly, Andrew; Batley, Richard (2018). Revisiting Consistency with Random Utility Maximisation: Theory and Implications for Practical Work. *Theory and Decision* 84 (2), 181-204.

Hidalgo-Ternero, Carlos Manuel (2020). Google Translate vs. DeepL: Analysing Neural Machine Translation Performance under the Challenge of Phraseological Variation. *MonTI Special Issue* 6, 154-177.

Hunger, Klaus (2003). *Industrial Dyes: Chemistry, Properties, Applications*. Weinheim: Wiley-VCH.

Hutchins, William John (2007). Machine Translation: A Concise History. *Computer-Aided Translation: Theory and Practice*, 13 (29-70), 1-21.

Hutchins, William John; Somers, Harold L. (1992). *An Introduction to Machine Translation.* London: Academic Press.

Kanoksilapatham, Budsaba (2003). *A Corpus-based Investigation of Scientific Research Articles: Linking Move Analysis and Multidimensional Analysis.* Dissertation, Georgetown University.

Kay, Martin (1982). Machine Translation. *American Journal of Computational Linguistics* 8 (2), 74-78.

Kelleher, John D. (2019). *Deep Learning*. Cambridge, MA: The MIT Press.

Kenny, Dorothy (2019). Machine Translation. In: Rawling, Piers; Wilson, Philip (eds). *The Routledge Handbook of Translation and Philosophy.* London: Routledge, 428-445.

Kirchhoff, Katrin; Capurro, Daniel; Turner, Anne (2012). Evaluating User Preferences in Machine Translation Using Conjoint Analysis. In: *Proceedings of the 6th Conference of European Association for Machine Translation (EAMT-12), Trento*, 119-126.

Kiritchenko, Svetlana; Mohammad, Saif M. (2017). Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. *Proceedings of the 55th*

Annual Meeting of the Association for Computational Linguistics (Short Papers), 465-470.

Koby, Geoffrey S.; Fields, Paul; Hague, Daryl; Lommel, Arle; Melby, Alan (2014). Defining Translation Quality. *Revista Tradumàtica: Tecnologies de la Traducció* 12, 413-420.

Koehn, Philipp (2010). *Statistical Machine Translation.* Cambridge: Cambridge University Press.

Koehn, Philipp (2020). *Neural Machine Translation.* Cambridge: Cambridge University Press.

Langer, Inghard; Schulz von Thun, Friedemann; Tausch, Reinhard (1974). *Verständlichkeit in Schule, Verwaltung, Politik, Wissenschaft: Mit einem Selbsttrainingsprogramm zur verständlichen Gestaltung von Lehr- und Informationstexten.* München/Basel: Reinhardt.

Läubli, Samuel; Sennrich, Rico; Volk, Martin (2018). Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4791-4796.

Levenshtein, Vladimir I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady* 10 (8), 707-710.

Lewin, Beverly A.; Fine, Jonathan; Young, Lynne (2001). *Expository Discourse: A Genre-based Approach to Social Science Research Texts.* London: Continuum.

Lewis, David M. (2014). Developments in the Chemistry of Reactive Dyes and their Application Processes. *Coloration Technology* 130 (6), 382-412.

LimeSurvey (2022). https://www.limesurvey.org/ (Accessed: 22 March 2022)

Lommel, Arle (2018). Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. In: Moorkens, Joss; Castilho, Sheila; Gaspari, Federico; Doherty, Stephen (eds). *Translation Quality Assessment: From Principles to Practice.* Cham: Springer International Publishing AG, 109-127.

Lommel, Arle; Burchardt, Aljoscha; Görög, Attila; Uszkoreit, Hans; Melby, Alan K. (2015). Multidimensional Quality Metrics (MQM) Issue Types. German Research Center for Artificial Intelligence. http://www.qt21.eu/mqm-definition/issues-list-2015-12-30.html (Accessed: 16 January 2022)

Louviere, Jordan J.; Flynn, Terry N.; Marley, Anthony A. J. (2015). *Best-Worst Scaling: Theory, Methods and Applications.* Cambridge: Cambridge University Press.

Luong, Minh-Thang (2016). *Neural Machine Translation*. Dissertation, Stanford University.

Macketanz, Vivien; Ai, Renlong; Burchardt, Aljoscha; Uszkoreit, Hans (2018). TQ-AutoTest – A Semi-Automatic Test Suite for (Machine) Translation Quality. In: Calzolari, Nicoletta; Choukri, Khalid; Cieri, Christopher; Declerck, Thierry; Goggi, Sara; Hasida, Koiti; Isahara, Hitoshi. Maegaard, Bente; Mariani, Joseph; Mazo, Hélène; Moreno, Asuncion; Odijk, Jan; Piperidis, Stelios; Tokunaga, Takenobu (eds). *Proceedings of the Eleventh International Conference on Language Resources and Evaluation. International Conference on Language Resources and Evaluation (LREC-2018) 11th May 7-12 Miyazaki Japan European Language Resources Association (ELRA) 2018*, 886-892.

Martínez Mateo, Roberto (2014). A Deeper Look into Metrics for Translation Quality Assessment (TQA): A Case Study. *Miscelánea: A Journal of English and American Studies* 49, 73-94.

Microsoft Bing Translator (2022). https://www.bing.com/translator (Accessed: 20 March 2022)

Montgomery, Scott L. (2009). English and Science: Realities and Issues for Translation in the Age of an Expanding Lingua Franca. *The Journal of Specialised Translation* 11, 6-16.

Munková, Daša; Wrede, Oľga; Absolon, Jakub (2019). Vergleich der menschlichen, maschinellen und Post-Editing-Übersetzung aus dem Slowakischen ins Deutsche mittels automatischer Evaluation. *Zeitschrift Für Slawistik* 64 (2), 231-261.

Neubig, Graham (2017). Neural Machine Translation and Sequence-to-sequence Models: A Tutorial. *arXiv preprint arXiv:1703.01619*.

Nießen, Sonja; Och, Franz Joseph; Leusch, Gregor; Ney, Hermann (2000). An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In: *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC´00). Athens: European Language Resources Association (ELRA)*, 39-45.

Nilsson, Nils J. (2010). *The Quest for Artificial Intelligence*. Cambridge: Cambridge University Press.

Nurminen, Mary (2019). Decision-Making, Risk, and Gist Machine Translation in the Work of Patent Professionals. *Proceedings of the 8th Workshop on Patent and Scientific Literature Translation*, 32-42.

Nwogu, Kevin Ngozi (1989). *Discourse Variation in Medical Texts: Schema, Theme and Cohesion in Professional and Journalistic Accounts.* Doctoral dissertation, University of Aston.

Olohan, Maeve (2007). The Status of Scientific Translation. *Journal of Translation Studies* 10 (1), 131-144.

Olohan, Maeve (2016). *Scientific and Technical Translation.* London and New York: Routledge.

Papineni, Kishore; Roukos, Salim; Ward, Todd; Zhu, Wei-Jing (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia*, 311-318.

Poibeau, Thierry (2017). *Machine Translation.* Cambridge, MA: The MIT Press.

Popel, Martin; Tomkova, Marketa; Tomek, Jakub; Kaiser, Łukasz; Uszkoreit, Jakob; Bojar, Ondřej; Žabokrtský, Zdeněk (2020). Transforming Machine Translation: A Deep Learning System Reaches News Translation Quality Comparable to Human Professionals. *Nature Communications* 11 (4381), 1-15.

Popović, Maja (2018). Error Classification and Analysis for Machine Translation Quality Assessment. In: Moorkens, Joss; Castilho, Sheila; Gaspari, Federico; Doherty, Stephen (eds). *Translation Quality Assessment: From Principles to Practice.* Cham: Springer International Publishing AG, 129-158.

Porsiel, Jörg (2020). Welcome to the Machine! Zwischen Goldgräberstimmung und der Suche nach dem Heiligen Gral. In: Porsiel, Jörg (ed.). *Maschinelle Übersetzung für Übersetzungsprofis.* Berlin: BDÜ Weiterbildungs- und Fachverlags GmbH, 2-10.

Rathenau Instituut (2020). Publication Output by Scientific Field, International Benchmark. https://www.rathenau.nl/en/science-figures/output/publications/publication-output-scientific-field-international-benchmark (Accessed: 12 December 2021)

SAE Mobilus (2001). Translation Quality Metric. Preview. https://saemobilus.sae.org/content/j2450_200112 (Accessed: 15 January 2022)

Schäffner, Christina (1997). From 'Good' to 'Functionally Appropriate': Assessing Translation Quality. *Current Issues in Language and Society* 4 (1), 1-5.

Schmalz, Antonia (2019). Maschinelle Übersetzung. In: Wittpahl, Volker (ed.). *Künstliche Intelligenz.* Berlin/Heidelberg: Springer, 194-211.

Sennrich, Rico; Haddow, Barry; Birch, Alexandra (2016). Neural Machine Translation of Rare Words with Subword Units. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany*, 1715-1725.

Shibata, Yusuke; Kida, Takuya; Fukamachi, Shuichi; Takeda, Masayuki; Shinohara, Ayumi; Shinohara, Takeshi; Arikawa, Setsuo (1999). Byte Pair Encoding: A Text Compression

Scheme that Accelerates Pattern Matching. *Technical Report DOI-TR-CS-161, Department of Informatics, Kyushu University*, 1-13.

Sismondo, Sergio (2010). *An Introduction to Science and Technology Studies.* Second edition. Chichester: Wiley Blackwell.

Snover, Matthew; Dorr, Bonnie; Schwartz, Richard; Micciulla, Linnea; Makhoul, John (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, Cambridge*, 223-231.

Somers, Harold L. (1999). Review Article: Example-Based Machine Translation. *Machine Translation* 14 (2), 113-157.

Specia, Lucia; Hajlaoui, Najeh; Hallett, Catalina; Aziz, Wilker (2011). Predicting Machine Translation Adequacy. *MTSUMMIT*, 513-520.

Statista (2020). Most Common Languages Used on the Internet as of January 2020, by Share of Internet Users. https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/#:~:text=Most%20common%20languages%20used%20on%20the%20internet%202020&text=As%20of%20January%202020%2C%20English,percent%20of%20global%20internet%20users. (Accessed: 12 December 2021)

Stein, Daniel (2009). Maschinelle Übersetzung – ein Überblick. *JLCL – Journal for Language Technology and Computational Linguistics* 24 (3), 5-18.

Stein, Daniel (2018). Machine Translation: Past, Present and Future. In: Rehm, Georg; Sasaki, Felix; Stein, Daniel; Witt, Andreas (eds). *Language Technologies for a Multilingual Europe.* Berlin: Language Science Press, 5-17.

Stymne, Sara; Ahrenberg, Lars (2012). On the Practice of Error Analysis for Machine Translation Evaluation. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul*, 1785-1790.

Surman, Jan (2016). Linguistic Precision and Scientific Accuracy: Searching for the Proper Name of "Oxygen" in French, Danish, and Polish. In: MacLeod, Miles; Sumillera, Rocío G.; Surman, Jan; Smirnova, Ekaterina (eds). *Language as a Scientific Tool: Shaping Scientific Language Across Time and National Tradition.* New York/London: Routledge, 131-148.

Swales, John M. (2004). *Research Genres: Explorations and Applications.* Cambridge: Cambridge University Press.

SYSTRAN Translate (2022). https://www.systran.net/en/translate/ (Accessed: 20 March 2022)

Tavosanis, Mirko (2019). Valutazione Umana di Google Traduttore e DeepL per le Traduzioni di Testi Giornalistici dall'Inglese verso l'Italiano. In: *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019). Bari, Italy. November 13-15, 2019.*

Vasheghani Farahani, Mehrdad (2020). Adequacy in Machine vs. Human Translation: A Comparative Study of English and Persian Languages. *Applied Linguisics Research Journal* 4 (5), 84-104.

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Łukasz; Polosukhin, Illia (2017). Attention Is All You Need. *31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, CA, USA*, 1-15.

Vilar, David; Xu, Jia; D'Haro, Luis Fernando; Ney, Hermann (2006). Error Analysis of Statistical Machine Translation Output. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006). Genoa*, 697-702.

Way, Andy (2018). Quality Expectations of Machine Translation. In: Moorkens, Joss; Castilho, Sheila; Gaspari, Federico; Doherty, Stephen (eds). *Translation Quality Assessment: From Principles to Practice.* Cham: Springer International Publishing AG, 159-178.

Wikipedia (2021). Transfer-Based Machine Translation. https://en.wikipedia.org/wiki/Transfer-based_machine_translation#/media/File:Direct_translation_and_transfer_translation_pyramid.svg (Accessed: 6 November 2021)

Wordpress (2014). Frederic Kaplan – L'Anglais comme Langue Pivot ou l'Impérialisme Linguistique Caché de Google Translate. https://fkaplan.wordpress.com/2014/11/15/langlais-comme-langue-pivot-ou-limperialisme-linguistique-cache-de-google-translate/ (Accessed: 12 December 2021)

Woyde, Rick (2001). Introduction to the SAE J2450 Translation Quality Metric. *Language International* 13 (2), 37-39.

Appendix 1

Source text passages used for empirical research

Source text passage 1:

> […] Phosphonate dyes could not hydrolyse during the application procedure to form the phosphonate ester bond with cellulose […]. It was important to include a dehydrating catalyst such as cyanamide or preferably dicyandiamide with the dye in order to promote the above esterification reaction. (Lewis 2014: 397)

Source text passage 2:

> When applied to nylon at pH 10, the dyes showed excellent substantivity, and yet at the boil the thiol-choline residue was eliminated to form the vinylsulphone dye which reacted readily with the deprotonated amines in nylon. (Lewis 2014: 404)

Source text passage 3:

> Selected acid 1:1 chrome complex dyes, because of their good wet- and lightfastness and their very good levelling power, and also 1:2 chrome complex and 1:2 cobalt complex dyes with hydrophilic groups, have successfully come into use as silk dyes. (Hunger 2003: 291)

Source text passage 4:

> With regard to heterocyclic compounds as coupling components, the importance of the formerly widespread pyrazolone-, aminopyrazole-, and 4-hydroxyquinolone-based yellow azo dyes has greatly diminished with the advent of the tinctorially superior and therefore more economical pyridone azo dyes. (Hunger 2003: 138)

Source text passage 5:

> These biological sorbents were found to be efficient in binding with basic dyes rather than acid dyes. This was attributed mainly to the Coulombic attraction between the negative surface of the adsorbent and the positively charged ions of basic dyes. (Al-Degs et al. 2000: 928)

Appendix 2

BIBD for the BWS in this thesis, featuring 5 sets with 4 translations each

Set 1	H	G	D	S
Source text passage	Phosphonate dyes could not hydrolyse during the application procedure to form the phosphonate ester bond with cellulose. It was important to include a dehydrating catalyst such as cyanamide or preferably dicyandiamide with the dye in order to promote the above esterification reaction.			
Human	Phosphonatfarbstoffe konnten während des Anwendungsverfahrens nicht zur Bildung von Phosphonatesterbindungen mit Zellulose hydrolysiert werden. Es war wichtig, dem Farbstoff einen dehydrierenden Katalysator wie Cyanamid oder vorzugsweise Dicyandiamid beizumischen, um die o. g. Veresterungsreaktion anzustoßen.			
Google Translate	Phosphonatfarbstoffe konnten während des Aufbringungsverfahrens nicht hydrolysieren, um die Phosphonatesterbindung mit Zellulose zu bilden. Es war wichtig, einen Dehydratisierungskatalysator wie Cyanamid oder vorzugsweise Dicyandiamid mit dem Farbstoff einzuschließen, um die obige Veresterungsreaktion zu fördern.			
DeepL Translator	Phosphonatfarbstoffe konnten während der Anwendung nicht hydrolysieren, um die Phosphonatesterbindung mit der Cellulose zu bilden. Es war wichtig, dem Farbstoff einen Dehydratisierungskatalysator wie Cyanamid oder vorzugsweise Dicyandiamid beizufügen, um die oben genannte Veresterungsreaktion zu fördern.			
SYSTRAN Translate	Phosphonatfarbstoffe konnten während des Applikationsvorgangs nicht zur Bildung der Phosphonatesterbindung mit Cellulose hydrolysieren. Es war wichtig, einen Dehydratisierungskatalysator wie Cyanamid oder vorzugsweise Dicyandiamid mit dem Farbstoff zu verbinden, um die obige Veresterungsreaktion zu fördern.			

Set 2	M	H	G	D
Source text passage	When applied to nylon at pH 10, the dyes showed excellent substantivity, and yet at the boil the thiol-choline residue was eliminated to form the vinylsulphone dye which reacted readily with the deprotonated amines in nylon.			
Microsoft Bing Translator	Wenn sie bei pH 10 auf Nylon aufgetragen wurden, zeigten die Farbstoffe eine ausgezeichnete Substantivität, und doch wurde beim Kochen der Thiol-Cholin-Rückstand eliminiert, um den Vinylsulfonfarbstoff zu bilden, der leicht mit den deprotonierten Aminen in Nylon reagierte.			
Human	Bei der Anwendung an Nylon bei pH-Wert 10 zeigten die Farbstoffe hervorragende Substantivität und trotzdem wurde beim Kochen der Thiol-Cholinrückstand eliminiert, sodass ein Vinylsulphonfarbstoff entstand, der leicht mit den deprotonierten Aminen im Nylon reagierte.			
Google Translate	Beim Auftragen auf Nylon bei pH 10 zeigten die Farbstoffe eine ausgezeichnete Substantivität, und doch wurde beim Sieden der Thiol-			

	Cholin-Rest eliminiert, um den Vinylsulfonfarbstoff zu bilden, der leicht mit den deprotonierten Aminen in Nylon reagierte.
DeepL Translator	Bei der Anwendung auf Nylon bei einem pH-Wert von 10 zeigten die Farbstoffe eine ausgezeichnete Substantivität, und dennoch wurde beim Kochen der Thiol-Cholin-Rest eliminiert, um den Vinylsulfon-Farbstoff zu bilden, der leicht mit den deprotonierten Aminen im Nylon reagierte.

Set 3	S	M	H	G
Source text passage	Selected acid 1:1 chrome complex dyes, because of their good wet- and lightfastness and their very good levelling power, and also 1:2 chrome complex and 1:2 cobalt complex dyes with hydrophilic groups, have successfully come into use as silk dyes.			
SYSTRAN Translate	Ausgewählte saure 1:1-Chromkomplexfarbstoffe sind aufgrund ihrer guten Nass- und Lichtechtheit und ihres sehr guten Verlaufsvermögens sowie 1:2-Chromkomplexfarbstoffe und 1:2-Kobaltkomplexfarbstoffe mit hydrophilen Gruppen erfolgreich als Seidenfarbstoffe in Einsatz gekommen.			
Microsoft Bing Translator	Ausgewählte Säure-1:1-Chromkomplexfarbstoffe sind aufgrund ihrer guten Nass- und Lichtechtheit und ihrer sehr guten Nivellierleistung sowie 1:2-Chromkomplex- und 1:2-Kobaltkomplexfarbstoffe mit hydrophilen Gruppen erfolgreich als Seidenfarbstoffe in Den Einsatz gekommen.			
Human	Erfolgreich als Seidenfarbstoffe eingesetzt werden ausgewählte saure 1:1-Chrom-Komplexfarbstoffe, aufgrund ihrer guten Feuchtigkeits- und Lichtbeständigkeit sowie der sehr guten nivellierenden Eigenschaften, sowie auch 1:2-Chrom- und 1:2-Cobalt-Komplexfarbstoffe mit hydrophilen Gruppen.			
Google Translate	Als Seidenfarbstoffe haben sich aufgrund ihrer guten Naß- und Lichtechtheit und ihres sehr guten Egalisiervermögens ausgewählte saure 1:1-Chromkomplexfarbstoffe sowie 1:2-Chromkomplex- und 1:2-Kobaltkomplexfarbstoffe mit hydrophilen Gruppen erfolgreich durchgesetzt.			

Set 4	D	S	M	H
Source text passage	With regard to heterocyclic compounds as coupling components, the importance of the formerly widespread pyrazolone-, aminopyrazole-, and 4-hydroxyquinolone-based yellow azo dyes has greatly diminished with the advent of the tinctorially superior and therefore more economical pyridone azo dyes.			
DeepL Translator	Bei den heterozyklischen Verbindungen als Kupplungskomponenten hat die Bedeutung der früher weit verbreiteten gelben Azofarbstoffe auf Pyrazolon-, Aminopyrazol- und 4-Hydroxychinolon-Basis mit dem Aufkommen der farblich überlegenen und damit kostengünstigeren Pyridon-Azofarbstoffe stark abgenommen.			
SYSTRAN Translate	Im Hinblick auf heterocyclische Verbindungen als Kupplungskomponenten hat sich die Bedeutung der bisher verbreiteten gelben Azo-Farbstoffe auf Pyrazolon-, Aminopyrazol- und 4-Hydroxychinolon-Basis mit dem Eintritt der farblich überlegenen und damit wirtschaftlicheren Pyridon-Azofarbstoffe stark verringert.			

Microsoft Bing Translator	Im Hinblick auf heterocyclische Verbindungen als Kopplungskomponenten hat die Bedeutung der ehemals weit verbreiteten Pyrazolon-, Aminopyrazol- und 4-Hydroxychinolon-basierten gelben Azofarbstoffe mit dem Aufkommen der tinktorial überlegenen und damit wirtschaftlicheren Pyridon-Azofarbstoffe stark abgenommen.
Human	Bezüglich heterocyclischer Verbindungen als Kupplungskomponenten ist die Bedeutung der vormals weit verbreiteten gelben Farbstoffe auf der Basis von Pyrazolon, Aminopyrazol und 4-Hydroxychinolin stark zurückgegangen, als die Pyridonazofarbstoffe aufkamen, mit ihrer überlegenen Färbekraft und somit höheren Sparsamkeit.

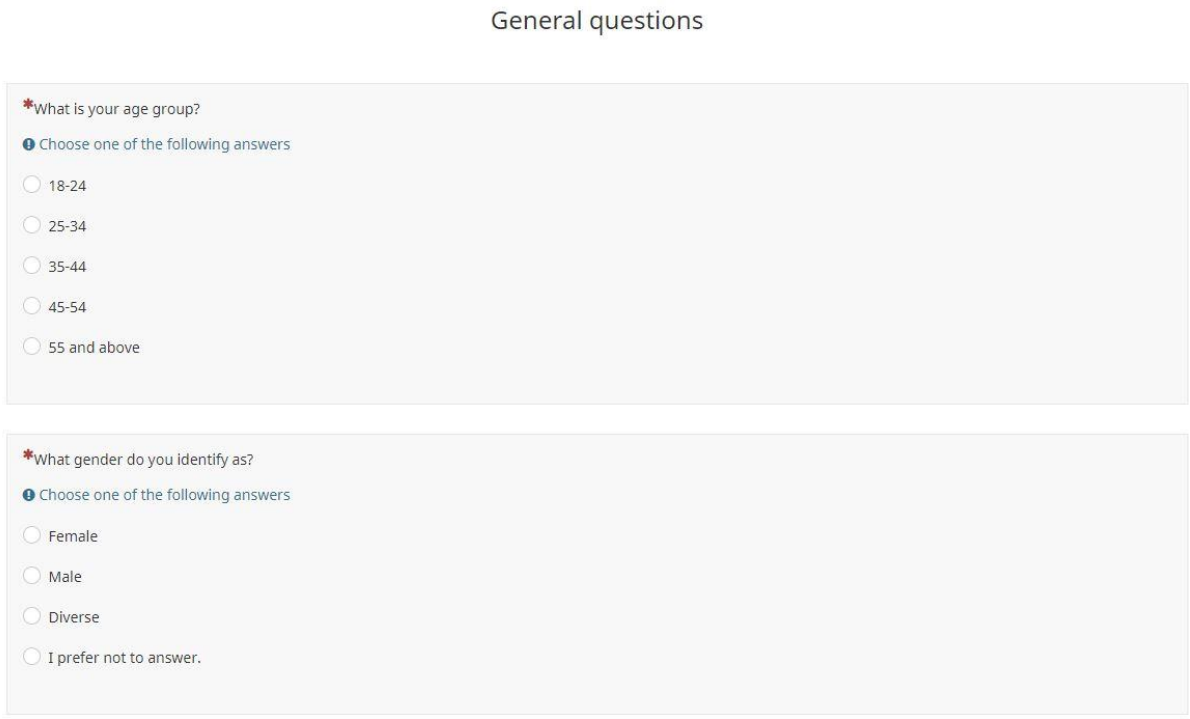
Set 5	G	D	S	M
Source text passages	These biological sorbents were found to be efficient in binding with basic dyes rather than acid dyes. This was attributed mainly to the Coulombic attraction between the negative surface of the adsorbent and the positively charged ions of basic dyes.			
Google Translate	Es wurde festgestellt, dass diese biologischen Sorptionsmittel bei der Bindung mit basischen Farbstoffen eher als mit sauren Farbstoffen wirksam sind. Dies wurde hauptsächlich der Coulomb-Anziehung zwischen der negativen Oberfläche des Adsorptionsmittels und den positiv geladenen Ionen basischer Farbstoffe zugeschrieben.			
DeepL Translator	Es wurde festgestellt, dass diese biologischen Sorptionsmittel eher basische Farbstoffe als saure Farbstoffe binden. Dies wurde hauptsächlich auf die Coulombsche Anziehungskraft zwischen der negativen Oberfläche des Adsorptionsmittels und den positiv geladenen Ionen der basischen Farbstoffe zurückgeführt.			
SYSTRAN Translate	Diese biologischen Sorbentien erwiesen sich als wirksam bei der Bindung mit basischen Farbstoffen anstelle von Säurefarbstoffen. Dies war hauptsächlich auf die Coulombische Anziehung zwischen der negativen Oberfläche des Adsorbens und den positiv geladenen Ionen basischer Farbstoffe zurückzuführen.			
Microsoft Bing Translator	Diese biologischen Sorbentien erwiesen sich als effizient bei der Bindung mit basischen Farbstoffen und nicht mit sauren Farbstoffen. Dies wurde hauptsächlich auf die Coulombische Anziehung zwischen der negativen Oberfläche des Adsorbens und den positiv geladenen Ionen der Basisfarbstoffe zurückgeführt.			

Appendix 3

Entire survey

1) Welcome page

2) General questions



*What is the highest educational degree you have completed?

Choose one of the following answers

- Bachelor's degree (BA, BSc, ...)
- Master's degree (MA, MSc, ...)
- Diploma (Mag.)
- Doctorate (Dr., PhD)
- Other (please specify in comment)

Please enter your comment here:

*In which field of study did you obtain that degree?

Choose one of the following answers

- Biology
- Chemistry
- Physics
- Translation studies
- Other (please specify in comment)

Please enter your comment here:

*How experienced are you with specialised language in the field of chemistry?

Choose one of the following answers

- 0 (no experience)
- 1
- 2
- 3
- 4 (very experienced)

Previous Next

3) Language skills

Language skills

*What is your first language?

Choose one of the following answers

- English
- German
- Other (please specify in comment)

Please enter your comment here:

*How would you rate your proficiency in **English** according to the Common European Framework of Reference for Languages (CEFR)?

Hover your mouse over each answer for more detailed information.

Choose one of the following answers

- C2: Proficient user - Mastery
- C1: Proficient user - Advanced
- B2: Independent user - Vantage
- B1: Independent user - Threshold
- A2: Basic user - Waystage
- A1: Basic user - Breakthrough

*How would you rate your proficiency in **German** according to the Common European Framework of Reference for Languages (CEFR)?

Hover your mouse over each answer for more detailed information.

Choose one of the following answers

- C2: Proficient user - Mastery
- C1: Proficient user - Advanced
- B2: Independent user - Vantage
- B1: Independent user - Threshold
- A2: Basic user - Waystage
- A1: Basic user - Breakthrough

4) How to complete

How to complete

Please complete the best-worst scaling in the next six questions according to this principle:

*You will be shown a set of four different German translations of the English text passage below.

Please indicate which of the four translations below you perceive to be the best and which one you perceive to be the worst in this set.

Please do not select "Best" and "Worst" for the same translation!

"English Text passage"

	Best	Worst
German translation 1	○	○
German translation 2	●	○
German translation 3	○	○
German translation 4	○	●

In this example, the participant perceived "German translation 2" as the best and "German translation 4" as the worst translation in this set.

Please note that it is not permitted to flag one and the same translation both as "best" and as "worst"!

Previous Next

5) Set 1

Set 1

*You will be shown a set of four different German translations of the English text passage below.

Please indicate which of the four translations below you perceive to be the **best** and which one you perceive to be the **worst** in this set.

Please do not select "Best" and "Worst" for the same translation!

"Phosphonate dyes could not hydrolyse during the application procedure to form the phosphonate ester bond with cellulose. It was important to include a dehydrating catalyst such as cyanamide or preferably dicyandiamide with the dye in order to promote the above esterification reaction."

	Best	Worst
Phosphonatfarbstoffe konnten während des Anwendungsverfahrens nicht zur Bildung von Phosphonatesterbindungen mit Zellulose hydrolysiert werden. Es war wichtig, dem Farbstoff einen dehydrierenden Katalysator wie Cyanamid oder vorzugsweise Dicyandiamid beizumischen, um die o. g. Veresterungsreaktion anzustoßen.	○	○
Phosphonatfarbstoffe konnten während des Aufbringungsverfahrens nicht hydrolysieren, um die Phosphonatesterbindung mit Zellulose zu bilden. Es war wichtig, einen Dehydratisierungskatalysator wie Cyanamid oder vorzugsweise Dicyandiamid mit dem Farbstoff einzuschließen, um die obige Veresterungsreaktion zu fördern.	○	○
Phosphonatfarbstoffe konnten während der Anwendung nicht hydrolysieren, um die Phosphonatesterbindung mit der Cellulose zu bilden. Es war wichtig, dem Farbstoff einen Dehydratisierungskatalysator wie Cyanamid oder vorzugsweise Dicyandiamid beizufügen, um die oben genannte Veresterungsreaktion zu fördern.	○	○
Phosphonatfarbstoffe konnten während des Applikationsvorgangs nicht zur Bildung der Phosphonatesterbindung mit Cellulose hydrolysieren. Es war wichtig, einen Dehydratisierungskatalysator wie Cyanamid oder vorzugsweise Dicyandiamid mit dem Farbstoff zu verbinden, um die obige Veresterungsreaktion zu fördern.	○	○

*Do you think the translation you have selected as the best one is of sufficient quality to be used in a professional context (e.g. in a research article or a scientific publication)? Feel free to comment on your choice.

Choose one of the following answers

○ Yes

○ No

Please enter your comment here:

*On a scale from **0 (lowest score)** to **4 (highest score)**, how good would you rate the translation you have just selected as the **best** one?

Choose one of the following answers

○ 4 (highest score)

○ 3

○ 2

○ 1

○ 0 (lowest score)

*On a scale from **-4 (lowest score)** to **0 (highest score)**, how bad would you rate the translation you have just selected as the **worst** one?

Choose one of the following answers

○ 0 (highest score)

○ -1

○ -2

○ -3

○ -4 (lowest score)

Previous

Next

6) Set 2

Set 2

*You will be shown a set of four different German translations of the English text passage below.

Please indicate which of the four translations below you perceive to be the **best** and which one you perceive to be the **worst** in this set.

Please do not select "Best" and "Worst" for the same translation!

"When applied to nylon at pH 10, the dyes showed excellent substantivity, and yet at the boil the thiol-choline residue was eliminated to form the vinylsulphone dye which reacted readily with the deprotonated amines in nylon."

	Best	Worst
Wenn sie bei pH 10 auf Nylon aufgetragen wurden, zeigten die Farbstoffe eine ausgezeichnete Substantivität, und doch wurde beim Kochen der Thiol-Cholin-Rückstand eliminiert, um den Vinylsulfonfarbstoff zu bilden, der leicht mit den deprotonierten Aminen in Nylon reagierte.	○	○
Bei der Anwendung an Nylon bei pH-Wert 10 zeigten die Farbstoffe hervorragende Substantivität und trotzdem wurde beim Kochen der Thiol-Cholinrückstand eliminiert, sodass ein Vinylsulphonfarbstoff entstand, der leicht mit den deprotonierten Aminen im Nylon reagierte.	○	○
Beim Auftragen auf Nylon bei pH 10 zeigten die Farbstoffe eine ausgezeichnete Substantivität, und doch wurde beim Sieden der Thiol-Cholin-Rest eliminiert, um den Vinylsulfonfarbstoff zu bilden, der leicht mit den deprotonierten Aminen in Nylon reagierte.	○	○
Bei der Anwendung auf Nylon bei einem pH-Wert von 10 zeigten die Farbstoffe eine ausgezeichnete Substantivität, und dennoch wurde beim Kochen der Thiol-Cholin-Rest eliminiert, um den Vinylsulfon-Farbstoff zu bilden, der leicht mit den deprotonierten Aminen im Nylon reagierte.	○	○

*Do you think the translation you have selected as the best one is of sufficient quality to be used in a professional context (e.g. in a research article or a scientific publication)? Feel free to comment on your choice.

Choose one of the following answers

○ Yes

○ No

Please enter your comment here:

*On a scale from **0 (lowest score)** to **4 (highest score)**, how good would you rate the translation you have just selected as the **best** one?

Choose one of the following answers

○ 4 (highest score)

○ 3

○ 2

○ 1

○ 0 (lowest score)

*On a scale from **-4 (lowest score)** to **0 (highest score)**, how bad would you rate the translation you have just selected as the **worst** one?

Choose one of the following answers

○ 0 (highest score)

○ -1

○ -2

○ -3

○ -4 (lowest score)

Previous

Next

7) Set 3

Set 3

*You will be shown a set of four different German translations of the English text passage below.

Please indicate which of the four translations below you perceive to be the **best** and which one you perceive to be the **worst** in this set.

Please do not select "Best" and "Worst" for the same translation!

"Selected acid 1:1 chrome complex dyes, because of their good wet- and lightfastness and their very good leveling power, and also 1:2 chrome complex and 1:2 cobalt complex dyes with hydrophilic groups, have successfully come into use as silk dyes."

	Best	Worst
Ausgewählte saure 1:1-Chromkomplexfarbstoffe sind aufgrund ihrer guten Nass- und Lichtechtheit und ihres sehr guten Verlaufsvermögens sowie 1:2-Chromkomplexfarbstoffe und 1:2-Kobaltkomplexfarbstoffe mit hydrophilen Gruppen erfolgreich als Seidenfarbstoffe in Einsatz gekommen.	○	○
Ausgewählte Säure-1:1-Chromkomplexfarbstoffe sind aufgrund ihrer guten Nass- und Lichtechtheit und ihrer sehr guten Nivellierleistung sowie 1:2-Chromkomplex- und 1:2-Kobaltkomplexfarbstoffe mit hydrophilen Gruppen erfolgreich als Seidenfarbstoffe in Den Einsatz gekommen.	○	○
Erfolgreich als Seidenfarbstoffe eingesetzt werden ausgewählte saure 1:1-Chrom-Komplexfarbstoffe, aufgrund ihrer guten Feuchtigkeits- und Lichtbeständigkeit sowie der sehr guten nivellierenden Eigenschaften, sowie auch 1:2-Chrom- und 1:2-Cobalt-Komplexfarbstoffe mit hydrophilen Gruppen.	○	○
Als Seidenfarbstoffe haben sich aufgrund ihrer guten Naß- und Lichtechtheit und ihres sehr guten Egalisiervermögens ausgewählte saure 1:1-Chromkomplexfarbstoffe sowie 1:2-Chromkomplex- und 1:2-Kobaltkomplexfarbstoffe mit hydrophilen Gruppen erfolgreich durchgesetzt.	○	○

*Do you think the translation you have selected as the best one is of sufficient quality to be used in a professional context (e.g. in a research article or a scientific publication)? Feel free to comment on your choice.

Choose one of the following answers

○ Yes

○ No

Please enter your comment here:

*On a scale from **0 (lowest score)** to **4 (highest score)**, how good would you rate the translation you have just selected as the **best** one?

Choose one of the following answers

○ 4 (highest score)

○ 3

○ 2

○ 1

○ 0 (lowest score)

*On a scale from **-4 (lowest score)** to **0 (highest score)**, how bad would you rate the translation you have just selected as the **worst** one?

Choose one of the following answers

○ 0 (highest score)

○ -1

○ -2

○ -3

○ -4 (lowest score)

Previous

Next

8) Set 4

Set 4

*You will be shown a set of four different German translations of the English text passage below.

Please indicate which of the four translations below you perceive to be the **best** and which one you perceive to be the **worst** in this set.

Please do not select "Best" and "Worst" for the same translation!

"With regard to heterocyclic compounds as coupling components, the importance of the formerly widespread pyrazolone-, aminopyrazole-, and 4-hydroxyquinolone-based yellow azo dyes has greatly diminished with the advent of the tinctorially superior and therefore more economical pyridone azo dyes."

	Best	Worst
Bei den heterozyklischen Verbindungen als Kupplungskomponenten hat die Bedeutung der früher weit verbreiteten gelben Azofarbstoffe auf Pyrazolon-, Aminopyrazol- und 4-Hydroxychinolon-Basis mit dem Aufkommen der farblich überlegenen und damit kostengünstigeren Pyridon-Azofarbstoffe stark abgenommen.	○	○
Im Hinblick auf heterocyclische Verbindungen als Kupplungskomponenten hat sich die Bedeutung der bisher verbreiteten gelben Azo-Farbstoffe auf Pyrazolon-, Aminopyrazol- und 4-Hydroxychinolon-Basis mit dem Eintritt der farblich überlegenen und damit wirtschaftlicheren Pyridon-Azofarbstoffe stark verringert.	○	○
Im Hinblick auf heterocyclische Verbindungen als Kopplungskomponenten hat die Bedeutung der ehemals weit verbreiteten Pyrazolon-, Aminopyrazol- und 4-Hydroxychinolon-basierten gelben Azofarbstoffe mit dem Aufkommen der tinktorial überlegenen und damit wirtschaftlicheren Pyridon-Azofarbstoffe stark abgenommen.	○	○
Bezüglich heterocyclischer Verbindungen als Kupplungskomponenten ist die Bedeutung der vormals weit verbreiteten gelben Farbstoffe auf der Basis von Pyrazolon, Aminopyrazol und 4-Hydroxychinolin stark zurückgegangen, als die Pyridonazofarbstoffe aufkamen, mit ihrer überlegenen Färbekraft und somit höheren Sparsamkeit.	○	○

*Do you think the translation you have selected as the best one is of sufficient quality to be used in a professional context (e.g. in a research article or a scientific publication)? Feel free to comment on your choice.

Choose one of the following answers

○ Yes

○ No

Please enter your comment here:

*On a scale from **0 (lowest score)** to **4 (highest score)**, how good would you rate the translation you have just selected as the **best** one?

Choose one of the following answers

○ 4 (highest score)

○ 3

○ 2

○ 1

○ 0 (lowest score)

*On a scale from **-4 (lowest score)** to **0 (highest score)**, how bad would you rate the translation you have just selected as the **worst** one?

Choose one of the following answers

○ 0 (highest score)

○ -1

○ -2

○ -3

○ -4 (lowest score)

Previous

Next

9) Set 5

Set 5

*You will be shown a set of four different German translations of the English text passage below.

Please indicate which of the four translations below you perceive to be the **best** and which one you perceive to be the **worst** in this set.

Please do not select "Best" and "Worst" for the same translation!

"These biological sorbents were found to be efficient in binding with basic dyes rather than acid dyes. This was attributed mainly to the Coulombic attraction between the negative surface of the adsorbent and the positively charged ions of basic dyes."

	Best	Worst
Es wurde festgestellt, dass diese biologischen Sorptionsmittel bei der Bindung mit basischen Farbstoffen eher als mit sauren Farbstoffen wirksam sind. Dies wurde hauptsächlich der Coulomb-Anziehung zwischen der negativen Oberfläche des Adsorptionsmittels und den positiv geladenen Ionen basischer Farbstoffe zugeschrieben.	○	○
Es wurde festgestellt, dass diese biologischen Sorptionsmittel eher basische Farbstoffe als saure Farbstoffe binden. Dies wurde hauptsächlich auf die Coulombsche Anziehungskraft zwischen der negativen Oberfläche des Adsorptionsmittels und den positiv geladenen Ionen der basischen Farbstoffe zurückgeführt.	○	○
Diese biologischen Sorbentien erwiesen sich als wirksam bei der Bindung mit basischen Farbstoffen anstelle von Säurefarbstoffen. Dies war hauptsächlich auf die Coulombische Anziehung zwischen der negativen Oberfläche des Adsorbens und den positiv geladenen Ionen basischer Farbstoffe zurückzuführen.	○	○
Diese biologischen Sorbentien erwiesen sich als effizient bei der Bindung mit basischen Farbstoffen und nicht mit sauren Farbstoffen. Dies wurde hauptsächlich auf die Coulombische Anziehung zwischen der negativen Oberfläche des Adsorbens und den positiv geladenen Ionen der Basisfarbstoffe zurückgeführt.	○	○

*Do you think the translation you have selected as the best one is of sufficient quality to be used in a professional context (e.g. in a research article or a scientific publication)? Feel free to comment on your choice.

Choose one of the following answers

○ Yes

○ No

Please enter your comment here:

*On a scale from **0 (lowest score)** to **4 (highest score)**, how good would you rate the translation you have just selected as the **best** one?

Choose one of the following answers

○ 4 (highest score)

○ 3

○ 2

○ 1

○ 0 (lowest score)

*On a scale from **-4 (lowest score)** to **0 (highest score)**, how bad would you rate the translation you have just selected as the **worst** one?

Choose one of the following answers

○ 0 (highest score)

○ -1

○ -2

○ -3

○ -4 (lowest score)

Previous

Next

10) Set 6

Set 6

*You will be shown a set of four different German translations of the English text passage below.

Please indicate which of the four translations below you perceive to be the **best** and which one you perceive to be the **worst** in this set.

Please do not select "Best" and "Worst" for the same translation!

"When applied to nylon at pH 10, the dyes showed excellent substantivity, and yet at the boil the thiol-choline residue was eliminated to form the vinylsulphone dye which reacted readily with the deprotonated amines in nylon."

	Best	Worst
Wenn sie bei pH 10 auf Nylon aufgetragen wurden, zeigten die Farbstoffe eine ausgezeichnete Substantivität, und doch wurde beim Kochen der Thiol-Cholin-Rückstand eliminiert, um den Vinylsulfonfarbstoff zu bilden, der leicht mit den deprotonierten Aminen in Nylon reagierte.	○	○
Bei der Anwendung an Nylon bei pH-Wert 10 zeigten die Farbstoffe hervorragende Substantivität und trotzdem wurde beim Kochen der Thiol-Cholinrückstand eliminiert, sodass ein Vinylsulphonfarbstoff entstand, der leicht mit den deprotonierten Aminen im Nylon reagierte.	○	○
Beim Auftragen auf Nylon bei pH 10 zeigten die Farbstoffe eine ausgezeichnete Substantivität, und doch wurde beim Sieden der Thiol-Cholin-Rest eliminiert, um den Vinylsulfonfarbstoff zu bilden, der leicht mit den deprotonierten Aminen in Nylon reagierte.	○	○
Bei der Anwendung auf Nylon bei einem pH-Wert von 10 zeigten die Farbstoffe eine ausgezeichnete Substantivität, und dennoch wurde beim Kochen der Thiol-Cholin-Rest eliminiert, um den Vinylsulfon-Farbstoff zu bilden, der leicht mit den deprotonierten Aminen im Nylon reagierte.	○	○

*Do you think the translation you have selected as the best one is of sufficient quality to be used in a professional context (e.g. in a research article or a scientific publication)? Feel free to comment on your choice.

Choose one of the following answers

○ Yes

○ No

Please enter your comment here:

*On a scale from **0 (lowest score)** to **4 (highest score)**, how good would you rate the translation you have just selected as the **best** one?

Choose one of the following answers

○ 4 (highest score)

○ 3

○ 2

○ 1

○ 0 (lowest score)

*On a scale from **-4 (lowest score)** to **0 (highest score)**, how bad would you rate the translation you have just selected as the **worst** one?

Choose one of the following answers

○ 0 (highest score)

○ -1

○ -2

○ -3

○ -4 (lowest score)

Previous

Next

11) Questions on machine translation

Questions on machine translation

*Have you ever used a machine translation tool before (e.g. DeepL, Google Translate, SYSTRAN Translate, Microsoft Bing Translator or similar)?

Choose one of the following answers

- ○ Yes
- ○ No

*How often do you use machine translation tools (e.g. DeepL, Google Translate, SYSTRAN Translate, Microsoft Bing Translator or similar)?

Choose one of the following answers

- ○ Never
- ○ Rarely
- ○ Sometimes
- ○ Often
- ○ Very often

For which purposes or in which situations do you use machine translation tools (e.g. DeepL, Google Translate, SYSTRAN Translate, Microsoft Bing Translator or similar)?

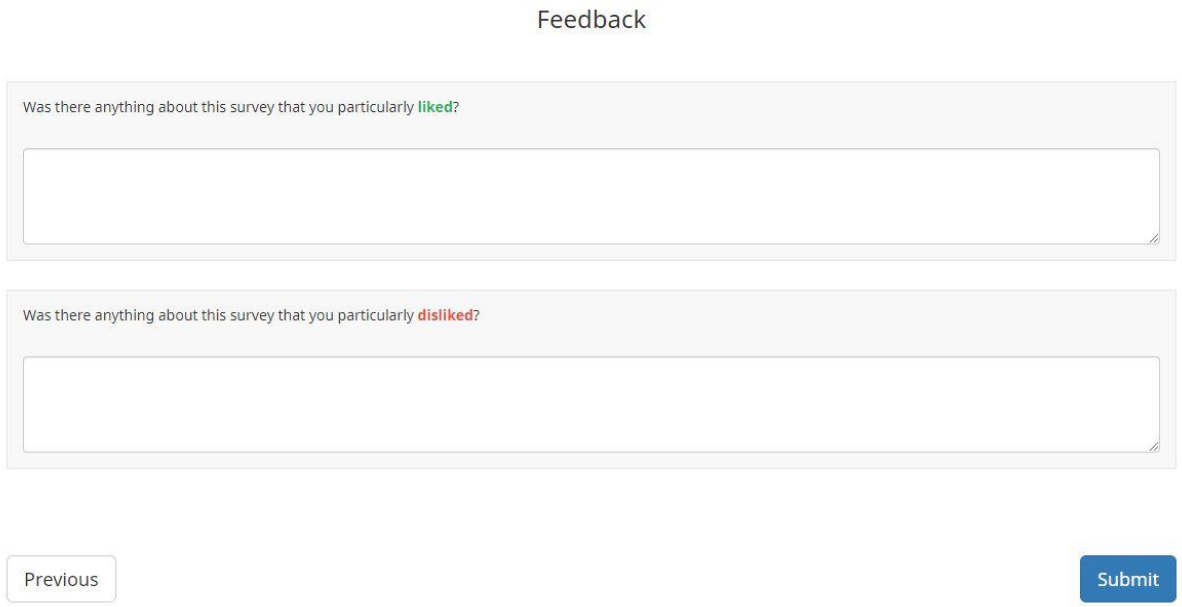
*When you use machine translation tools, how satisfied are you usually with the results?

Choose one of the following answers

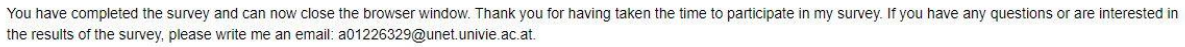
- ○ Very dissatisfied
- ○ Somewhat dissatisfied
- ○ Neutral
- ○ Somewhat satisfied
- ○ Very satisfied

Previous Next

12) Feedback



13) Closing page



Abstract

English:

The objective of this master's thesis is to determine how translators and natural scientists rate the raw output of neural machine translation in comparison to human translation in highly specialised texts from the domain of chemistry and in the language direction English to German, and to determine which of the neural machine translation tools used in this experiment (DeepL Translator, Google Translate, Microsoft Bing Translator, and SYSTRAN Translate) yields the best results without any human intervention such as pre- or post-editing.

Using a best-worst scaling, the translations produced by the neural machine translation tools and the human translation were compared and assessed by the participants of the survey. By means of plus and minus points, participants could indicate how good or bad they considered the respective translations to be, resulting in a numeric score for each translation method.

The findings of this survey were that both the translators and the natural scientists rated one neural machine translation tool, namely DeepL Translator, significantly better than all the remaining objects, even better than human translation. DeepL Translator was also the only neural machine translation tool with a positive numeric rating. In conclusion, it can be said that, under certain conditions, neural machine translation tools can outperform human translation without undergoing post-editing, although this cannot be claimed for all tools.

Deutsch:

Ziel dieser Masterarbeit ist es, zu ermitteln, wie Übersetzer*innen und Naturwissenschaftler*innen den unbearbeiteten Output neuronaler maschineller Übersetzung im Vergleich zur menschlichen Übersetzung bei hochspezialisierten Texten aus dem Bereich der Chemie und in der Sprachrichtung Englisch-Deutsch bewerten und welches der in diesem Experiment verwendeten neuronalen maschinellen Übersetzungstools (DeepL Translator, Google Translate, Microsoft Bing Translator und SYSTRAN Translate) ohne menschliche Eingriffe wie Pre- oder Post-Editing die besten Ergebnisse liefert.

Anhand eines Best-Worst-Scalings wurden die von den neuronalen maschinellen Übersetzungstools erstellten Übersetzungen und die menschliche Übersetzung verglichen und von den Teilnehmer*innen der Umfrage bewertet. Durch Plus- und Minuspunkte konnten die Teilnehmer*innen angeben, für wie gut oder schlecht sie die jeweiligen Übersetzungen hielten, wodurch eine numerische Punktezahl für jede Übersetzungsmethode erlangt wurde.

Das Ergebnis dieser Umfrage war, dass sowohl die Übersetzer*innen als auch die Naturwissenschaftler*innen ein neuronales maschinelles Übersetzungstool, nämlich DeepL Translator, deutlich besser bewerteten als alle anderen Objekte, sogar besser als die menschliche Übersetzung. DeepL Translator war auch das einzige neuronale maschinelle Übersetzungstool mit einer positiven numerischen Bewertung. Zusammenfassend lässt sich sagen, dass neuronale maschinelle Übersetzungstools ohne Post-Editing unter bestimmten Bedingungen die menschliche Übersetzung übertreffen können, obwohl dies nicht für alle Tools behauptet werden kann.