universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Quality Effects of Domain Adaptation of NMT Models with Translation Memories in the Occupational Health and Safety Domain“

verfasst von / submitted by

Giacomo Munda 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 070 348 331
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Translation Italienisch Deutsch
Betreut von / Supervisor:	Ass.-Prof. Mag. Dr. Dagmar Gromann, BSc

Acknowledgments

First of all, I would like to sincerely thank Professor Gromann for mentoring me with passion, patience and incredible expertise in every part of this journey.

Another thank you goes to the colleagues at Eurac Research, Dr. Elena Chiocchetti and Dr. Flavia De Camillis, without whom this project would not have been possible. Thank you for trusting me and constantly sharing with me your amazing experience and knowledge in the language industry. Especially Elena deserves a special thanks, I will tremendously miss her endless patience, expertise and motivation that she was able to impart to me during the last year.

To my friends in Italy, Tommaso and Gianmarco, even if do not tell them enough, they played a big role in my life during the last years.

Finally, to my family, who supported me and believed in me. Especially to my mom and dad for their constant presence and their endless love.

Table of contents

1. Introduction *7*

1.1. Research question and assumptions 8

1.2. Motivation and objectives 9

2. Theoretical background *11*

2.1. A brief history of machine translation 11

2.2. Rule-Based Machine Translation (RBMT) 14

2.3. Statistical Machine Translation (SMT) 15
2.3.1. Word-based approach 15
2.3.2. Phrase-based approach 17

2.4. Neural Machine Translation (NMT) 20
2.4.1. Artificial neural networks 20
2.4.2. Word embeddings and the encoder-decoder architecture 22
2.4.3. Corpora and translation memories for NMT training 23
2.4.3.1. Data preprocessing and cleaning 24
2.4.4. Domain adaptation for NMT 26

2.5. Quality evaluation for MT 30
2.5.1. Automatic evaluation 30
2.5.2. Human evaluation 32

2.6. German as a pluricentric language 33
2.6.1. South Tyrolean German 33
2.6.1.1. Historical background 34
2.6.1.2. Developing a unique terminology 35
2.6.1.3. Occupational health and safety in South Tyrol 36
2.6.1.4. South Tyrolean German as a low-resource language 39

3. Related work on domain adaptation *41*

4. Methodology *44*

4.1. Domain selection 44

4.2. Experimental settings 45

4.3. Training and test data 46
4.3.1. Preprocessing the data 48
4.3.2. Dataset splitting 50

4.4. Domain adaptation 51
4.4.1. Selecting a pre-trained model 51
4.4.2. ModernMT 52

4.5. Quality evaluation 54
4.5.1. Automatic evaluation – BLEU and METEOR 54
4.5.2. Human evaluation – taxonomy 54

5. Results *58*

5.1. Results of the preprocessing steps 58

5.2. Results of the dataset splitting 59
5.3. Results of the automatic evaluation 61
5.4. Results of the human evaluation 63
6. Discussion *74*
7. Conclusions *77*
8. Bibliography *79*
9. Appendix *89*
9.1. Source text 89
9.2. Translation performed by the adapted model 93
9.3. Translation performed by the baseline model 97
Abstract *103*
Abstract in German *104*

List of figures

Figure 1: example of a phrase-based model (Koehn 2010: 128) 18
Figure 2: an artificial neural network with two hidden layers 21
Figure 3: fine-tuning in a CAT scenario (Koehn 2020: 256) 28
Figure 4: instance-based NMT adaptation (Farajian et al. 2018) 28
Figure 5: Südtirol genormt-label in the bistro entry for the term ‘questura’ 36
Figure 6: accuracy errors and mistranslations annotated during the evaluation 63
Figure 7: fluency errors annotated during the evaluation 64

List of equations

Equation 1: probability function 16
Equation 2: results of the lexical translation probability distribution function 17

List of tables

Table 1: example-phrases for BLEU 31
Table 2: data-overview for the experiment 48
Table 3: type of errors to annotate during the evaluation 56
Table 4: noise-cleaning operations on the SSL and TU corpora 59
Table 5: results of the dataset splitting operations 60
Table 6: overall quality improvement in terms of BLEU and METEOR scores 61
Table 7: translation example 1 65
Table 8: translation example 2 66
Table 9: translation example 3 67
Table 10: translation example 4 68
Table 11: translation example 5 69
Table 12: translation example 6 70
Table 13: translation example 7 71
Table 14: translation example 8 72
Table 15: translation example 9 72

1. Introduction

In today's language industry landscape, the image of the classic translator working with printed dictionaries, writing translations on paper has disappeared. Translators nowadays, as do many other job profiles and industries, rely on technology to produce high quality target texts. It is common in the translation process to utilize Computer Assisted Translation (CAT) tools, but in the last decades Machine Translation (MT) has gained a lot of momentum and even many professional translators acknowledge its benefits and its improvement (Zaretskaya et al. 2015; Macken et al. 2020; Vieira 2020; Way 2020). Due to globalization, the amount of translations required worldwide has increased drastically and it is impossible for humans to complete all of those in a timely manner. While MT was considered a promising option to accelerate the translation process, a devastating report on its progress and results in the 1960s, the ALPAC report (ALPAC 1966), discontinued funding endeavors in the US and several other countries. Nevertheless, after a few years, the first successful MT system was introduced and research started to grow once again. The exponential growth of computing power allowed to create very sophisticated systems which are used today. After the success of Statistical Machine Translation (SMT), today the state-of-the-art approach is Neural Machine Translation (NMT). Even though NMT has reached unexpected quality levels, especially regarding fluency, there are still several issues of NMT results that should be scientifically addressed. Beside the very common problems of not being always able to adapt to different contexts or mistranslating words with multiple meanings, other domain-specific issues can be observed in MT models. The common MT systems available on the web are usually not trained with specific data of a particular domain only because they have to perform well on a general basis. When NMT systems are faced with out-of-domain sentences, the results are often very poor and the quality of the output depends very much on the training conditions (Koehn & Knowles 2017). Therefore, companies that need to translate domain-specific texts usually need to train the system or rely on systems that have previously been trained with in-domain data in order to get a quality improvement (Way 2018; Way 2020). Although MT has been evolving for more than 50 years now, the problem of high-resource and low-resource language pairs still exists and trained models perform very well with language pairs containing English and worse with less-spoken languages. In addition, specific domains, such as the legal domain, especially connected to language varieties, represent a challenge for modern MT approaches (Heiss & Soffritti 2018; Wiesmann 2019; De Camillis 2021).

The language pair that will be used in this research is Italian and South Tyrolean German. The latter differs from common German mostly from a legal point of view. Another challenge of MT comes when translating legal-administrative texts. Every legal system is different and concepts may seem very similar, but even a small difference can cause misunderstandings. If a system does not contain a perfect match of a legal concept, it is very likely that the result will be incorrect. Furthermore, the specific language variety of South Tyrolean German, especially its legal and administrative terminology, is very peculiar. With South Tyrol being an official bilingual province in Italy, both German and Italian must be bound to the Italian legal system. Consequently, German terminology must be adapted to represent equivalent concepts pertaining to a different legal system. Several terms can be borrowed from other German-speaking countries, but the majority of the terms utilized today were established by a Terminology Commission in 1988 (Chiocchetti 2019). This peculiarity can represent a major issue for an MT system. Fortunately, the South Tyrolean research center *Eurac Research* agreed to provide roughly 25,000 translation units on the specific domain of occupational health and safety to this research endeavor, which will be utilized to investigate the quality effects of domain adaptation on a pre-trained NMT model. The majority of the translation units are e-learning modules that Eurac Research translated and revised for the Autonomous Province of Bolzano and distributed to employees and companies in South Tyrol with the purpose of educating employees on occupational health and safety and replacing several courses held face-to-face. The remaining translation units come from the aligned version of the Consolidated Law on Safety at Work. For the experimental part of this thesis, an NMT engine called *ModernMT* (Bertoldi et al. 2018) that allows to integrate translation memories for a real-time domain adaptation of the system will be utilized.

1.1. Research question and assumptions

Having identified the problems of MT that are relevant for this research, namely the difficulties of translating into German from the Italian legal system, translating specific terminology of occupational health and safety correctly and adapting to a specific domain in a low-resource language setting, it is necessary to formulate a research question on which the master's thesis will focus. Specifically, the present work aims at answering the following question: *How and to what extent will the quality of an NMT system improve with domain*

adaptation based on domain-specific translation memories on occupational health and safety in the language pair Italian – South Tyrolean German?

To address this question, two parallel corpora are used as input for the adaptive NMT system ModernMT. In this way, the parallel corpora should allow ModernMT to adapt to the domain of occupational health and safety before automatic translation. The translation memories originally contained roughly 25,000 in-domain segments but are further cleaned and filtered. The texts produced by the baseline and the adapted models are evaluated automatically with BLEU and METEOR. A more detailed human evaluation follows, using a taxonomy to identify major error categories.

Based on previous research (Servan et al. 2016), 25,000 translation units should be sufficient to adapt a pre-trained model to a domain and to consequently increase automatic evaluation scores by adapting terminology, style etc. Furthermore, the method of instance-based adaptation used by ModernMT showed valid effectiveness in translating domain specific terminology (Farajian et al. 2018) and should allow the model to rapidly adapt to a specific domain by providing only a relatively small amount of data.

1.2. Motivation and objectives

There are many arguments that motivate this research. Increasing the quality of an NMT system is of course one of the main objectives of researchers in this field. The results could contribute to estimating the size of data required for domain adaptation within companies, especially for low-resource languages such as South Tyrolean German, since several companies already perform domain adaptation, e.g. utilizing MT engines incorporated in CAT tools. Furthermore, the approach is targeted at people with little knowledge of computer science. With only a few steps, they could customize their own NMT model without retraining it from scratch and without diving into complex fine-tuning techniques, which would require deep technical expertise in the field of natural language processing and machine/deep learning. The adaptation techniques used by the chosen system are complex and require a technical understanding, but their usage on the commercial product is user-friendly and suitable for freelance translators and smaller companies, which do not necessarily possess highly technical competences. Specifically, one of the main users could potentially be public administration offices, in the case of this research the Autonomous Province of Bolzano. There have been little attempts to create an adapted MT system aiming at translating legal-administrative texts in South Tyrol, despite the continuous need for new

translations. To the best of my knowledge, Contarino (2021) produced the first and only research on this topic by adapting a system with a parallel corpus retrieved from *LexBrowser* (n.d.), a publicly available bilingual collection of South Tyrolean legislation. For the present research endeavor, the data that will be used to adapt a pre-trained model are partly publicly available resources, such as the translation of the Consolidated Law on Safety at Work, which the South Tyrolean public administration offices can easily retrieve. Other data are internal translation memories. Furthermore, new training courses on occupational health and safety are currently written, translated and shared with companies and employees in South Tyrol (Eurac Research 2022a; Eurac Research 2022b). Therefore, using an MT system that is trained for that specific domain would be very beneficial to the province as well as research on low-resource domain adaptation.

2. Theoretical background

This chapter is dedicated to the theoretical background of the present master's thesis. It starts with a detailed description of machine translation, covering its history and the main approaches developed to reach the state-of-the-art neural machine translation used today. After a description of the main architectures, other important components of neural machine translation are explained, such as parallel corpora, common procedures of preprocessing and cleaning the data before using them as input for the machine, the approaches utilized for the domain adaptation of NMT models as well the quality evaluation techniques for MT. Since the language pair used in the experimental part of this work contains the low-resource language of South Tyrolean German, an analysis of this language variety follows, together with a description of the occupational health and safety domain in the context of South Tyrol and an explanation of why this German variety can be considered low-resourced.

2.1. A brief history of machine translation

Training and adapting an NMT system require some theoretical knowledge and it is vital to know how the different approaches proposed until now work. Machine translation was invented in 1933, but one of the pioneers of this technology is without no doubt Warren Weaver, who developed the mathematical model behind the first computer-based machine translation system (Weaver 1952). He looked at machine translation as an encoding-decoding activity, where a foreign language is written in different symbols and the computer has the task of decoding it. Thanks to the invention of new and sophisticated computers, researchers were hopeful that the MT-problem was soon going to be solved. Nevertheless, ambiguity and several other language problems were not sufficiently considered. The Automatic Language Processing Advisory Committee (ALPAC) decided to commission a study to analyze the progress and status of MT. The ALPAC report had catastrophic consequences in the research field of MT, since it came to the conclusion that costs for post-editing were still too high and no real advantage in using MT was to be seen (Koehn 2010). Even though funding was stopped temporarily in the US after the publication of the ALPAC report, in the following decades, the first successful MT systems were introduced, which were building on the method of Rule-Based Machine Translation (RBMT). These systems used explicit rules to generate the source text and even though they were not able to produce high quality target texts compared to the results of the current approaches, mainly because

of ambiguity problems and issues regarding word order etc., they remained available for several years. Examples of the first RBMT systems are: Météo (Thouin 1981), SYSTRAN (Toma 1977), Logos (Scott 1989) and Metal (Bennet & Slocum 1985)

In the 1990s, another important milestone for translation technology was reached, namely computer assisted translation tools, which will play an important role in this master's thesis as well. These systems allowed to store all translations produced by human translators in a database in form of Translation Memories (TMs), which helps the translator with consistency and productivity by not having to re-translate similar sentences. Whenever a match is found, this will automatically be proposed to the translator, who will analyze it and decide if it is suitable in the context or not (Sandrini 2012). Translation Memory Systems (TMS) spread the idea that "it may be better to learn how to translate from past translation examples" (Koehn 2020: 36). This approach was soon taken into consideration to develop new kinds of MT systems that no longer needed to follow linguistic rules. They are called data-driven systems and the input to generate the translation are millions of parallel sentences that form a corpus. In the 1990s, IBM (Brown et al. 1993) developed the first successful statistical machine translation systems, which used statistical likelihood to produce target texts. After the relatively unsuccessful word-based approach, phrase-based SMT (Marcu & Wong 2002; Koehn et al. 2003) slowly became the state-of-the-art for more than a decade and researchers kept improving it by introducing new methods and approaches. The problem of SMT was the need of large parallel corpora that were not always easily available. During the years, with the growth of the internet and the continuous research in MT, digital text resources became more available, which contributed to the improvement of these types of systems. Furthermore, CAT became an essential tool for professional translators, who were able to produce new data in form of aligned translation units that served as input to SMT as well. In addition, MT was integrated into TMS to facilitate and accelerate the translation process. In the following decades, the largest software companies such as Google, IBM, Microsoft, etc. released their own commercial SMT systems and open-source toolkits to develop and train customized SMT models were introduced. The most famous one was *Moses* (Koehn et al. 2007). Machine Translation was no longer a topic of interest for developers and computer scientists only. Everyone with access to the world wide web could make use of it. Nevertheless, adding new features to SMT led to a stagnation in the development of these systems, since they required much human maintenance and the current computing power started to not be enough to keep up with the trend.

A new technology was needed. Research on neural networks and on possible machine translation approaches based on neural computation already started long before the first NMT systems became available. Already in the 1980s and 1990s, researchers investigated this new technology, but without any useful results since they could test the new approaches with very limited data sizes and iterations due to hardware limitations and neural networks required much more computing power than available at the time. After almost two decades, neural methods were integrated into statistical machine translation, but there were still many concerns, especially regarding the use of Graphic Processing Units (GPUs), a powerful alternative to Central Processing Units (CPUs) used to train neural networks. In the 2010s, models aiming at producing pure neural machine translation were introduced (Kalchbrenner & Blunsom 2013; Cho et al. 2014; Sutskever et al. 2014) and after a few years and some adjustments, NMT became the state-of-the-art approach by outperforming SMT in almost every task. The reason behind this success was the completely new technology on which NMT is based. So-called Machine Learning (ML) and Deep Learning (DL) use Artificial Neural Networks (ANNs), which learn the translation task by themselves after having been trained with large parallel corpora. ANNs can therefore look at a whole sentence instead of phrases and produce highly fluent target texts. Unfortunately, unlike previous MT approaches, it is very difficult to control and manage the output generated by the machine. Training and especially fine-tuning already trained NMT models plays therefore a crucial role to improve the translation performance and possibly adapt the models to specific domains as will be done in this master's thesis. In order to provide a solid basis for the experimental part of this thesis, an analysis of the three main approaches to MT, especially NMT, will follow in the next chapters.

2.2. Rule-Based Machine Translation (RBMT)

This is considered the classical approach of machine translation and it uses morphological, syntactic and semantic information about the languages that are being translated in form of explicit linguistic rules (Shiwen & Xiaojing 2014). This information is retrieved from monolingual, bilingual or multilingual dictionaries and grammars. RBMT systems generate the target text based on the information that is gathered during the analysis process. There are three main approaches used in these types of systems, namely the direct translation, the transfer-based translation and the interlingua translation.

The **direct translation** does not need to pass through an additional representation (Tripathi & Sarkhel 2010) and the input is just translated into another language by looking up words in a dictionary and by following some handwritten rules on morphology, syntax, word order etc. The first step is to use a dictionary to find direct equivalences of the words that need to be translated. After that, a common rule of RBMT systems is to make use of the alignment function, utilized for example to reorder nouns and adjectives (e.g *macchina rossa* in Italian, but *rotes Auto* in German) (Poibeau 2017). The **transfer-based translation** splits the process into three steps: first a linguistic *analysis* of the source text is needed, then the syntactic and semantic structures of the source language are transferred into the same structures of the target language during the *transfer* process and finally, during the *synthesis* process, the constituents in the source language are replaced by the target language equivalents (Tripathi & Sarkhel 2010: 390). In the last approach, the **interlingua translation**, the source language text is analyzed into an abstract universal language-independent representation (interlingua) and the target language is generated from the interlingua by "using the lexical units and the syntactic constructions of the target language" (Alansary 2014: 42). Nevertheless, this last approach has never reached its expected potential due to the difficulty of devising an effective interlingua.

2.3. Statistical Machine Translation (SMT)

RBMT was considered the first working machine translation approach that was able to efficiently translate a text from one language to another. Nevertheless, using dictionaries, grammars and other handwritten rules could not contribute to solving the translation problem, since language was far more complex than a simple encoding-decoding activity. As mentioned in Section 2.1, IBM (Brown et al. 1993) developed the first SMT models: a revolutionary approach to MT to which this chapter will be dedicated. Nevertheless, this new technology took a while to surpass the previous RBMT, especially the interlingua approaches, and gained momentum at the beginning of the new century when the IBM algorithms became widely available. Furthermore, Koehn (2010: 18) points out that the terrorist attacks of September 11, 2001, played a big role for a renewed interest in the automated translation of foreign languages, especially Arabic. Substantial funding was provided for this field and the growth of the internet, together with the rapid increase of computing power and storage contributed to the success of SMT.

MT was no longer based on linguistic rules, the source text would be generated by learning the language from parallel corpora used to train the systems and by following statistical and probability models. Knowing the theory behind SMT is very beneficial to understanding the future concept of neural machine translation. The following subchapters are based mostly on the excellent book of Koehn (2010).

2.3.1. Word-based approach

The very first SMT systems were based on the so-called word-based approach, which was soon replaced by the more sophisticated and efficient phrase-based approach. Nevertheless, it still introduces many key points of SMT in general, which are vital for a deeper understanding of this technology.

Word-Based Statistical Machine Translation (WBSMT) is based on lexical translation, which can be considered as the translation of isolated words, as seen previously with the direct and transfer-based translation in RBMT. However, the difference between these two technologies lays in the input from which the machine learns how to translate words. WBSMT does not use fixed bilingual dictionaries, but a large corpus where all possible translations can be found. In these parallel corpora it is common to find different translations

for different words. Koehn (2010: 81) gives an example of how the German word *Haus* could be translated:

Haus – house, building, home, household, shell

In most cases, it is likely that the English translation *house* would be correct, but there are some specific circumstances in which other translations are possible and this is the reason why statistics can be used to find the most suitable variant. For example, by analyzing a large German-English parallel corpus and by counting the occurrences of the translation of the word *Haus*, it is possible to learn how and how often the source word has been translated. Hypothetically, for 10,000 occurrences of the word *Haus*, 8,000 times it has been translated as *house*, 1,600 times as *building*, etc.

Given these statistics, the question is how the models can select the correct translation from numerous possibilities. Human translators usually look at the context and/or are able to understand different domains. WBSMT uses the Equation 1 called **lexical translation probability distribution**:

$$p_f : e \rightarrow p_f(e)$$

Equation 1: probability function

In the example above, where *f* is the foreign word *Haus*, the function "returns a probability, for each choice of English translation *e*, that indicates how likely that translation is" (Koehn 2010: 82). If a translation is very common, the result should be a high value. If it is less common, the function should return a lower value, which cannot be 0, otherwise the translation is impossible. The values must range between 0 and 1 and the sum of all probabilities $p_f(e)$ must be 1. According to Koehn (2010: 82), this type of estimation obtained from the function is called **maximum likelihood estimation**, since it estimates the likelihood of the data. By applying the function to the example of the translation of *Haus*, the results represented in Equation 2 can be obtained:

$$p_f(e) = \begin{cases} 0.8 & \text{if } e = house \\ 0.16 & \text{if } e = building \\ 0.02 & \text{if } e = home \\ 0.015 & \text{if } e = household \\ 0.005 & \text{if } e = shell \end{cases}$$

Equation 2: results of the lexical translation probability distribution function

Having obtained the necessary values to decide which translation is the most suitable in the context, it is vital to rely on an alignment function to reorder the single words, since languages often have different word orders that have to be taken into consideration, and on language and translation models to successfully translate sentences into another language.

Word-based statistical machine translation models were able to partially solve some of the problems of the automated translation process. Promising concepts such as deficiency, i.e., handling the problem of the possibility for an output word to be inserted into a position that is already occupied by another word and fertility, i.e., a function $n(\emptyset|f)$ that indicates, for each word *f* in the source sentence, the possible number $\emptyset$ of target words, which can be 0 if the source word is dropped in the target language, were used. Nevertheless, this approach quickly led to several difficulties when confronted with more complex sentences like idioms (which in most cases have completely different constructs in the target language), prepositions and words whose meaning shifts across domains. That is the reason why phrase-based machine translation was developed at the beginning of the 21st century. Words as atomic units were no longer the focus of interest, instead the goal of this new approach was to create new models for translating small word sequences at a time.

2.3.2. Phrase-based approach

Translating single words and then reordering them through multiple alignment models is not very beneficial. Even though the word-based approach was an extraordinary revolution in the field of MT, several other language problems could not be solved. Considering for example the German word *kitzeklein*, word-based models were able to assign 0, 1 or *n* words in English. The correct translation in this case would be *very small*. Conversely, if two words in the source language should correspond to one word in the target language, the system would not be able to translate the words correctly. This is of course a serious deficiency of the word-based approach since multiword expressions often occur in the translation of natural languages. There are some methods which are able to improve the previous word-

based models, but they often require large data to perform additional alignment steps and they are not considered the most efficient solution. Another way of looking at this problem is to translate and align at a phrase level (i.e., sequences of words), which would enable to better take the context into account and thus produce higher quality translations. This is the main idea behind Phrase-Based Statistical Machine Translation (PBSMT), a revolutionary approach that remained the state-of-the-art for more than a decade. Fig. 1 illustrates in a simplified way how the phrase-based approach works.

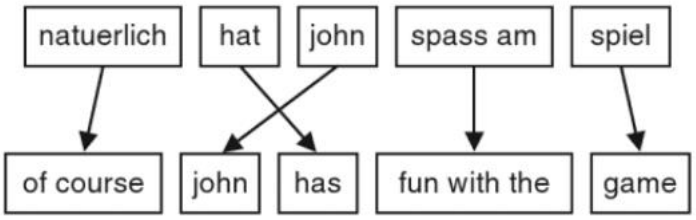

Figure 1: example of a phrase-based model (Koehn 2010: 128)

Instead of assigning every source word to a target word and relying on the alignment models to reorder words, PBSMT simplifies the process and divides the sentences into multiple phrases. By adopting this approach, the number of elements that need to be aligned clearly decreases and multiword expressions are possible in both directions. Some phrases still need to be reordered, as can be seen in the example in Fig. 1, the English verb must follow the subject. In order to calculate the most probable translation of phrases, the lexical translation probability distribution function no longer works and it must be adapted to generate a probabilistic phrase translation table. The translation model thus contains "a set of target-language phrases, each of which is associated with a list of source-language translations for those phrases and corresponding probabilities" (Hearne & Way 2011: 212).

The data are of course retrieved from large parallel corpora, which became exponentially larger and allowed to produce meaningful translation probabilities. However, although being able to solve many language problems, the phrase-based models failed to take into consideration distant dependencies and to achieve an effective reordering on a global level (Koehn et al. 2003). Syntax-based machine translation successfully addressed these issues and took advantage of parsers to syntactically analyze the input to determine its correct structure. However, the models needed much more data than the phrase-based

models to function correctly, resulting in slow performances as well as high memory requirements (Williams et al. 2016).

SMT, especially phrase-based SMT, was the state-of-the-art approach for many years and the most famous systems such as Google Translate, SYSTRAN etc. relied on this technology to produce high quality translations. Nevertheless, research on Artificial Neural Networks (ANN) was prospering. In 2015, the first neural machine translation systems were proposed and one year later this new technology started outperforming SMT in almost all language pairs, becoming the new state-of-the-art still used today (Koehn 2020).

2.4. Neural Machine Translation (NMT)

Neural machine translation, from a very broad perspective, can be considered similar to statistical machine translation, in the sense that they are both data-driven systems. These systems are trained with large parallel corpora with usually aligned sentences of already existing translations, such as translation memories. The number of sentences that serve as input to an NMT system varies from hundreds of thousands to even millions and the training may require several days, weeks or even months. These two technologies differ, however, in the computational approach, because NMT builds on artificial neural networks.

2.4.1. Artificial neural networks

The primary goal of artificial neural networks is to mimic the connection patterns that exists in the human brain when actual neurons activate themselves to pass an input signal from one cell to another, hence the name. The main difference lays in the way of how they process and learn information. As Koehn (2020: 31) points out, natural neurons evolve by themselves without being given the right answer when fully trained. One option to train artificial neurons is a supervised training process and a weight adjustment after having been given the right answer from a large amount of data.

Artificial neural networks were introduced long before the first neural machine translation systems. Already in 1957, the first neural network models were introduced by Rosenblatt (1957), who called them **perceptrons**. Unlike modern neural networks, they only had an input layer and an output layer and researchers were very confident about this new technology. However, it became evident that the training was computationally very demanding and specific hardware, such as Graphic Processing Units (GPUs) used today, was not powerful enough at the time (Forcada 2017). After a few years, with the introduction of new and more sophisticated hardware, ANNs with more layers, the so-called hidden layers, were introduced as exemplified in Fig. 2.

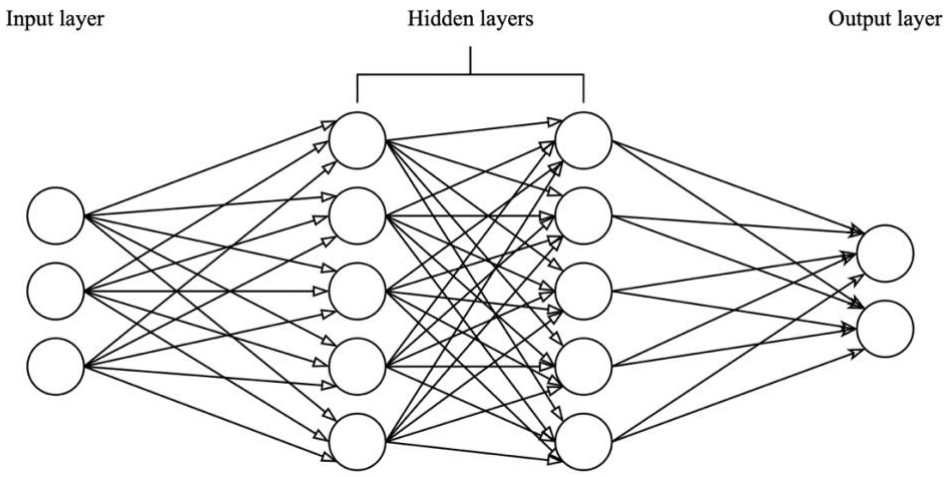

Figure 2: an artificial neural network with two hidden layers

In simple terms, every ANN is organized in multiple layers. The input layer transmits the information to the output layer by passing through a pre-defined number of hidden layers. The computation that happens in these layers is not visible to the user and that is the reason why the training process of neural networks is often described as a black box (Aggarwal 2018). Every layer consists of several nodes and the connections between these nodes initially carry random weights. The weighted sum of the inputs is calculated, an additional input as bias is added and the result is fed to an activation function, which is responsible for the activation of only those nodes that crosses the threshold value. The value is then passed to other nodes of the network and the operation is repeated until it reaches the output layer. These types of neural networks are also called **feedforward networks**, since the flow of information is transmitted to the next layers in a forward direction, but the real power and benefit of ANNs come from another algorithm of supervised training called **backpropagation** (Rumelhart et al. 1986). Backpropagation, short for 'backward propagation of error', plays a very important role in neural network's training and aims at optimizing weight values by propagating back the information containing the error to earlier hidden layers (Koehn 2020). Having reliable training data available is crucial because through this process training examples are fed into the network and the output generated by the neural model is compared to the output of the training example. The weights are adjusted and updated through several iterations until the output corresponds to the target translations in the training dataset. This process is usually repeated several times and each pass over the entire dataset is called an **epoch**.

During the years, many other ANN architectures, such as the Recurrent Neural Networks (RNN), the Convolutional Neural Networks (CNN), the Long Short-Term Memory (LSTM) architecture and the current state-of-the-art Transformer were invented (Tan et al. 2020; Koehn 2020). These types of neural networks build the foundations of neural machine translation and they constitute one of the most essential parts of neural machine translation models.

2.4.2. Word embeddings and the encoder-decoder architecture

One crucial question is the ability of neural networks to represent words, phrases, and language per se. According to Koehn (2020: 104), using token IDs or bit encodings for token IDs would be useless since words with similar encodings may have completely different meanings. That is the reason why the **one-hot encoding** was developed. This method generates vectors containing a list of binary numbers (0, 1), where only the 1 in a certain position decodes the word or the value. The length of the numbers in the vector corresponds to the number of words in the vocabulary. However, one-hot encoding fails at capturing relations and similarities between words and cannot be used for translation tasks. Thus, in modern systems, they are utilized as input vectors to neural networks when trained on natural language. The actual semantic representation of words learned by neural networks are instead dense numeric vectors that are the weights on the hidden layers, called **word embeddings** (Mikolov et al. 2013). The benefit of word embeddings comes from using decimals between 0 and 1, since the weights are finally passed through a *softmax* activation function that translates weights to a realm between 0 and 1, which allows to represent words in similar contexts and predict the next sequences of words. Word embeddings are usually trained with monolingual corpora input to a feedforward neural network and every vector represents the contextual meanings of a specific word so that words that are closer in the so-called vector space are more likely to have a similar meaning.

Early NMT systems were models based on an encoder-decoder architecture (Sutskever et al. 2014; Cho et al. 2014), also called sequence-to-sequence (seq2seq). The encoder has the task of mapping a source language sentence into vector embeddings of individual words and the decoder decodes the numerical representations one word at a time into a target-language sentence by calculating the maximum likelihood of each word (Forcada 2017: 297-298). This approach was soon improved by introducing the so-called **attention mechanism** (Bahdanau et al. 2014), which completely revolutionized NMT approaches. The attention

mechanism eschews the use of Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) and compared to the seq2seq models "the decoder pays attention (responds) not only to the last representation built by the encoder […] but also to the whole sequence of representations built during encoding […] through an appropriate additional set of neural connections and layers" (Forcada 2017: 299). By doing so, it becomes possible to access information about every element of the input sequence. Nevertheless, this approach still processes each element of the sequence one at a time and is subjected to computational inefficiency when dealing with very large corpora. The current state-of-the-art architecture is the so-called **Transformer** (Vaswani et al. 2017). These models are entirely based on attention mechanisms, which Vaswani et al. (2017) call **self-attention** and are able to extract word-features to extrapolate the importance of all words in a given sentence and putting the spotlight on the most important words. Furthermore, they generate context-sensitive word representations that, unlike previous models, depend on the entire source sentence and not on single words and can therefore improve quality and fluency (Stahlberg 2020).

2.4.3. Corpora and translation memories for NMT training

The main difference between rule-based and data-driven approaches in MT is the input required to train the systems. Supervised training procedures for NMT, as a data-driven approach, rely on parallel corpora to produce meaningful translations and as Zoph et al. (2016) point out, this new technology is considered particularly data-hungry. Therefore, in order to build reliable NMT models, there is a strong need to train them with substantial amounts of parallel sentences to achieve better results than previous approaches such as SMT. The learning curve is steeper compared to SMT and an insufficient quantity of data would lead to worse performances (Koehn & Knowles 2017). This requirement makes it harder to train NMT systems on low-resource languages, since the availability of parallel corpora for languages with a smaller community of speakers, e.g. South Tyrolean German, is often scarce (see Section 2.6.1.4 for a closer look at this low-resource language variety).

There are several ways to produce parallel corpora. In the language industry landscape, parallel corpora are usually created by translators themselves in form of parallel segments called translation memories, which are produced when translating new texts. Translation memories are usually saved in the so-called Translation Memory eXchange (TMX) format, which allows to exchange these databases between any tool that supports the format (Melby & Wright 2015). Another way of producing parallel corpora is to align a monolingual corpus

with its translation. The downside of this method is that the corpus (especially if retrieved from web sources) must undergo a cleaning and filtering process before being utilized to train custom MT systems. Creating parallel corpora in a CAT environment has the benefit that the translator decides which units to save into the TM database by discarding segments containing noisy elements and thus producing more reliable data. Furthermore, duplicate sentence pairs, which could be harmful to the system, are automatically discarded by the TMS, resulting in less redundant data. Today, large collections of data that can be utilized to train custom MT models are available online, many of which can be retrieved on OPUS (n.d.) and, depending on the language pair, they can contain up to billions of tokens.

2.4.3.1. Data preprocessing and cleaning

The abundance of data retrievable online has grown exponentially in the last decades and since machine learning is deployed in many different tasks, the need for large quantities of data has become essential. Data nowadays, especially if retrieved online, can contain several noisy elements, which turned out to negatively affect the quality of NMT systems. Furthermore, a problem of web-scraped parallel corpora is translations that have been produced by MT systems and although the quality of MT has substantially improved, they are not the type of input adequate to train new systems with (Rarrick et al. 2011).

Text preprocessing consists in "converting a raw text file […] into a well-defined sequence of linguistically meaningful units" (Palmer 2010: 9) so that machine learning algorithms can work with high quality data and thus perform better. Broadly speaking, text preprocessing consists in three important steps: tokenization, normalization and noise removal. For the practical part of this work, only the noise removal part will be adopted, but it is vital to know why all preprocessing steps are important when creating a machine learning model. **Tokenization**, in natural language processing, is one of the basic text processing steps that consists in splitting a raw text into smaller pieces, called *tokens* (Koehn 2010: 34). A paragraph can be tokenized into sentences and a sentence can be tokenized into words. The result is an explicitly demarcated sequence of words that can be further processed by a computer. The next step is **normalization**, a set of tasks to obtain uniformed representations in a text, e.g. lowercasing or truecasing words (i.e., preserving the uppercase in names), removing punctuation, lemmatizing words (i.e., the process of transforming words to their base form, e.g. *doing* to *do*, *car's* to *car*), converting digits to words etc. It is important to state that all preprocessing steps are adopted before training or adapting a

system and once the machine has been trained, processes such as **detokenization** must be adopted to reattach punctuation and undo other tokenization and normalization steps in order to generate a fluid text with correct capitalization, punctuation etc. (Koehn 2010: 35). The third step, **noise removal**, consists in removing all those elements in a text that can deteriorate the quality of MT systems. Especially NMT was found to be particularly sensitive to noise in corpora. Compared to statistical machine translation, where adding more data (regardless if noisy or not) resulted in an increase of automatic evaluation scores, adding new data to NMT models could lead to catastrophic consequences if not preprocessed correctly (Koehn 2020). In order to find out which noise elements most deteriorate the quality of an NMT system, Khayrallah and Koehn (2018) carried out an experiment in which they deliberately polluted a parallel corpus with noisy elements. According to their findings, typical categories of noise in parallel corpora include:

- **Untranslated sentences**: the source segment is the same of the target segment and no translation is to be seen. This particular trait is often seen in web-scraped corpora and according to Khayrallah and Koehn (2018: 78), it leads the machine to learn copying behaviors, which will further apply to other non-polluted segments resulting in a substantial quality-loss of up to 24.0 BLEU points.
- **Misaligned sentences**: the target segment does not match the source segment. This issue can easily occur when automatically aligning corpora from the web. To solve this problem, sentence pairs with a different token or character length (Kurfali & Östling 2019; Hangya & Fraser 2018) or sentences with a higher length ratio than a given threshold can be discarded (Gupta et al. 2019).
- **Misordered words**: translations produced by MT, low quality human translations or highly specialized language can cause disfluent language that might affect the quality of the model (Khayrallah & Koehn 2018).
- **Wrong language**: a corpus, beside the correct language pair, can contain a third language that can lower the overall quality since the engine is not trained to deal with additional languages, besides the ones used for training (Khayrallah & Koehn 2018).
- **Long and short segments**: according to Gupta et al. (2019), sentences with a smaller or bigger length ratio than the average can be discarded. Nevertheless, this noise type does not affect dramatically the overall quality of the system and short segments consisting of 3-5 tokens can even improve the BLEU score, whereas very short segments (≤ 2 tokens) only lower the score by maximum 1.0 BLEU point (Khayrallah & Koehn 2018)

- **Non-linguistic characters**: segments composed mostly by nonlinguistic characters such as punctuation, tags, whitespaces, digits etc. are considered useless and can be discarded (Gupta et al. 2019; Barbu et al. 2016). Segments can be discarded if 50% of the content consists of non-alphabetical symbols or one of the sides contains more non-alphabetical symbols than the other side (Rikters 2018).
- **Duplicate sentence pairs**: translation units which repeats themselves are usually discarded. If working with CAT tools, this process is usually already carried out automatically by TMS, resulting in cleaned translation memories. Depending on the task, almost duplicate sentence pairs can be discarded to lower the number of redundant data in the corpus (Pinnis 2018).

The previous noise categories serve as examples to identify which elements most deteriorate the quality of NMT systems, but filtering processes are task-specific and must be adopted according to many factors: the origin of the retrieved parallel corpus, the source of translation, the domain of the parallel corpus and the availability of data for the training or adaptation procedure. These characteristics strongly influence preprocessing steps and a first manual analysis of the data is crucial to understand future cleaning and filtering steps. Filtering a parallel corpus might lead to a substantial loss of data needed for training, it is therefore necessary to wisely choose which elements, according to their influence on the translation quality, need to be filtered out.

2.4.4. Domain adaptation for NMT

Parallel corpora used to train (neural) machine translation systems are usually collections of human translations that cover a broad range of topics. Their purpose is to create a system that knows how to translate natural language without focusing on a specific domain. Data-driven systems, especially NMT, allows a specialization of MT models to particular domains, which in the case of this master's thesis is the occupational health and safety domain, to get better translation performances. This specialization is called **domain adaptation** and it differs from a common training procedure, because researchers/developers etc. usually rely on already trained systems, or perform a training from scratch first, and then try to specialize the models by adding domain-specific parallel data. As Koehn (2020: 243) succinctly points out, the question is why a translation model should not consist of only in-domain data. Broadly speaking, the amount of in-domain data is usually small and it could

happen that several words, which need to be translated, do not occur at all in the training data. In such cases, the system would not be able to translate the words correctly, resulting in incomplete sentences. Out-of-domain data serve exactly this purpose and since their volume is much bigger, their task is to fill the gaps where the in-domain data cannot be deployed. Many methods for performing domain adaptation are utilized to increase translation performances. The so-called **data interpolation** method consists in duplicating the in-domain data so that the training algorithm encounters them more often. The **model interpolation** method consists instead in creating two models with in-domain and out-of-domain data respectively and ensembling them to create a unified model with all the necessary data. The most popular method for domain adaptation in NMT is called **fine-tuning** and consists in training a model with in-domain and out-of-domain data and then performing additional iterations on the in-domain data only. One of the main downsides of this method is the possible degradation of the translation quality when faced with out-of-domain sentences due to the small amount of in-domain data used for training (Freitag & Al-Onaizan 2016).

Domain adaptation has already reached the practical world of the language industry, where MT is being utilized in CAT tools such as Translation Memory Systems (TMS). While the translator works on a document, the MT engine incorporated in the TMS suggests its solutions. If the translator modifies the output generated by the machine, the new data will serve as input to proceed with a constant domain adaptation so that the second time a word that has previously been corrected by the translator appears to the machine, it should be able the generate a correct translation. This circular adaptation is exemplified in Fig. 3. It builds on the method of fine-tuning and is effective even with small amounts of in-domain data (Servan et al. 2016). In detail, Servan et al. (2016), in addition to a generic model utilized to train the network, created five models, every one of them containing a different amount of in-domain data. The first improvement occurred already after adding 500 lines, but the first substantial improvement could be seen by adding 5,000 lines of in-domain data. They obtained an increase of the BLEU score of more than 2.5 points and after adding the whole document (922,000 lines) the BLEU score improved by 15 points.

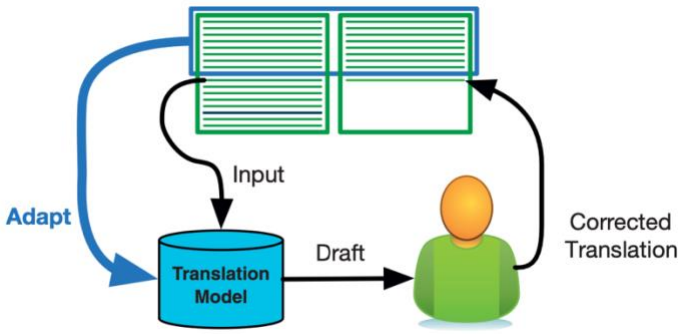

Figure 3: fine-tuning in a CAT scenario (Koehn 2020: 256)

Another interesting method is called **instance-based adaptation**. It was developed by Farajian et al. (2017) and it plays an important role in this master's thesis since it is currently deployed in the commercial NMT system ModernMT. Instance-based adaptation works as following: given a baseline NMT model, a parallel corpus and a sentence to be translated, the latter is used to retrieve a set of parallel segments from the corpus that are similar to the sentence that needs to be translated. If a good match is found, the retrieved parallel segments are used to adapt the hyperparameters of the NMT model and the resulting locally-tuned model is then applied to translate the given sentence (Farajian et al. 2017: 129). If no similar sentences in the corpus are found, the baseline model will translate the sentence. This process is exemplified in Fig. 4.

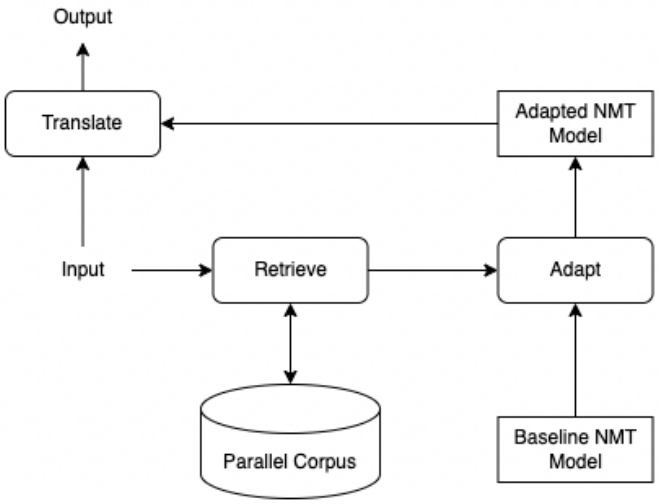

Figure 4: instance-based NMT adaptation (Farajian et al. 2018)

Other adaptation methods focus on applying terminology constraints to the domain adaptation process (Sato et al. 2020). In some cases, rendering specific terms correctly is considered more important than producing fluent translations, as could be the case of companies which need their systems to translate brand-specific terms correctly. Terminology can be contained in the classical in-domain data and most of the time, domain-specific data are sufficient to reach the goal of adapting a system, but researchers developed techniques to augment the traditional MT input as parallel sentences with terminological databases to produce more reliable models (Dinu et al. 2019, Hasler et al. 2018).

2.5. Quality evaluation for MT

Over the years, machine translation has substantially improved and NMT has reached incredible quality levels that in some cases exceed the work of human translators, especially regarding fluency. Nevertheless, target texts often need to be improved in a post-editing environment. Methods for training and improving MT systems require reliable methods for evaluating their quality. The output of an MT system can be assessed through an automatic or a human evaluation, whereby the latter is surely more reliable than the former but far more time-consuming and less objective, since different evaluators may have completely opposite opinions regarding the accuracy of a translation. Automatic metrics require much less time and resources but the majority of them need translations created by humans as a reference and their quality is often assumed and not verified (Rivera-Trigueros 2021).

2.5.1. Automatic evaluation

MT evaluation is an essential and complex task, which contributes to the continuous growth and improvement of these kinds of systems. Automatic metrics utilized for assessing MT quality have been introduced to facilitate this demanding process and compared to a human evaluation they drastically reduce the time invested in this activity, the overall costs by reducing human intervention, they are more objective and consistent (Popović 2018). Furthermore, they can be utilized repeatedly while performing experiments on MT to assess preliminary results. As just mentioned, the majority of these metrics compare the output generated by the machine with one or more reference translations ideally created by experienced translators and the result is a numerical output. Due to the growing research on MT, several metrics have been created over the years, but the most popular and the *de facto* standard is the Bilingual Evaluation Understudy (BLEU) score developed by Papineni et al. (2002). According to Marie et al. (2021), BLEU is used in 98.8% of the papers focusing on MT experiments and the central idea behind this algorithm and many other evaluation metrics is the following: “The closer a machine translation is to a professional human translation, the better it is” (Papineni et al. 2002: 311). BLEU analyzes the precision of the match by comparing *n-grams* (sequences of *n* words) of the candidates. A perfect match, i.e., all the tokens into which a sentence has been split correspond exactly to the ones of the reference translation, would result in the number 1. Conversely, if none of the tokens are alike, it would result in a perfect mismatch and therefore, the output generated by the metric would be 0. In many publications, for better readability, the result is multiplied by 100.

BLEU, as many other automatic metrics, are very literal and differences in the sentence length as well as the usage of synonyms penalize the score. Therefore, a valid translation might be considered incorrect if relying only on automatic metrics, such as BLEU.

Reference: This is a small test for calculating the BLEU score.
Candidate: This is a test to calculate the BLEU score.

Table 1: example-phrases for BLEU

For instance, the actual BLEU score of the example in Table 1 would be 0.7, due to the omission of 'small' and the different tense of the verb 'calculate'.

Other metrics that take synonyms and other morphological variants into account have been developed following BLEU. An example is METEOR, developed by Banerjee and Lavie (2005). METEOR was designed to solve the already mentioned weaknesses of BLEU and compared to the latter it is based on a combination of unigram matching, unigram precision and unigram recall between the candidate translation and the reference translation. *Unigrams*, compared to n-grams that split a sentence into sets of one, two or more words, divide a sentence into single word sequences and allow a word-to-word matching, in which the matching supports not only identical words, but also synonyms and words that have and identical stem (Banarjee & Lavie 2005: 66). The algorithm, after an alignment-phase between two strings based on the mapping of unigrams, computes a score that takes into consideration more reference translations as well. In this case, the candidate translation created by the MT system is compared to all available references and the best score is reported (Banarjee & Lavie 2005: 67). Nevertheless, even though METEOR allows a more detailed syntactical and semantical analysis, this type of evaluation remains a challenging task for automatic metrics (Castilho et al. 2018).

Popular metrics which are not based on n-grams but calculate the number of edits required to match the reference translation include Translation Error Rate (TER) and Word Error Rate (WER). TER (Snover et al. 2006) determines the amount of post-editing required on a translation produced by a machine. Compared to metrics such as BLEU or METEOR, a lower score means that the MT model performed well and less post-editing is needed. WER (Nießen et al. 2000) computes the number of edits in terms of insertions, deletions and substitutions needed to match the reference translation. It was first developed to evaluate the quality of speech recognition models.

2.5.2. Human evaluation

A human evaluation requires more steps and a simple numerical output to assess complex linguistic phenomena is not sufficient. Evaluating translation quality has always been a controversial topic, mostly because different translation projects require different evaluation methods (Lommel et al. 2014). To assess this problem, Lommel at al. (2014) developed the so-called Multidimensional Quality Metrics (MQM), a framework which allows users to customize their own quality metrics based on a blueprint containing 20 of the most common issues that arise when evaluating translation quality. These issues are enclosed in seven major categories: accuracy, fluency, design, locale convention, style, terminology and verity. MQM was designed to be adopted in the evaluation of human- and machine-produced texts. Human assessment of MT is often performed in terms of the high-level categories of *fluency* and *accuracy* (Castilho et al. 2018; Koehn 2010) and by looking at reference translations or at the original text, the revisor evaluates the semantic quality (accuracy) in order to find out if the information of the source text has been transferred appropriately (Rivera-Trigueros 2021: 5). Fluency focusses instead on the target text and is defined as "the extent to which the translation follows the rules and norms of the target language (regardless of the source or input text)" (Castilho et al. 2018: 18). Other quality evaluation metrics can be performed via ranking by comparing several machine translations of the same text and defining which version is the best and which one is the worst (Callison-Burch et al. 2007). Nevertheless, according to Popović (2018), MT researchers find themselves wanting to have additional information about the systems they develop by understanding where the machine lacks in quality performance, which are the strengths and weaknesses of the system and what type of modifications allow for an improved engine. To this end, error classification and analysis have been introduced. Depending on the assessment goal, several taxonomies can be adopted, but the main error categories regarding accuracy include addition, omission, mistranslation, disambiguation, errors regarding gender, inconsistent punctuation between source text and target text and terminology. Typical fluency errors include grammar, lexicon, orthography and coherence.

2.6. German as a pluricentric language

German, as many other languages such as English or Spanish, is considered a pluricentric language with an official status in multiple countries distributed across the globe. The term *pluricentric* refers to a language that is used in several nations as an official language and therefore does not have identical linguistic features but provides a national variety with its own norms (Kloss 1978). An important feature is that for each variety there is a reference work, e.g. a dictionary that codifies it. A peculiarity of such languages is the fact that they unify people with different nationalities through the language but at the same time "they separate them through the development of national norms and indices and linguistic variables with which the speakers identify" (Clyne 2012: 1). These linguistic varieties, although different in many levels, still represent the standard language (e.g. German, English, Spanish etc.)

For instance, German has an official status in seven European countries: Germany, Austria, Switzerland, Luxembourg, Belgium, Italy and Liechtenstein. The German varieties present in Germany, Austria and Switzerland are considered fully official languages and they differ from each other through the existence of a series of instances that establish the norms of the standard German of the specific country, such as austriacisms, helvetisms and teutonisms. Such differences are particularly evident in the lexicon and in prosodic or phonetic features, but they play a relevant role in other linguistic characteristics as well, such as in the morphosyntax or in the pragmatics.

Regarding the German variety present in Italy, it is spoken in the Autonomous Province of Bolzano by roughly 300,000 people (ASTAT 2021) and is commonly referred to as South Tyrolean German. Compared to the three official German varieties, South Tyrolean German is not a fully-fledged variety and it is only used from a relatively small community in the northernmost province of Italy. The next subsections are dedicated to this particular language variety and to why its characteristics play an important role in this master's thesis.

2.6.1. South Tyrolean German

Today, South Tyrol can be considered a region with high democratic principles admired by many other multilingual areas in Europe. Nevertheless, the presence of three different linguistic groups (i.e., Italian, German and Ladin) with equal rights, regardless of their language, is the result of a complicated journey, which led to the peaceful coexistence of

these three distinct groups (Ralli & Stanizzi 2018). The development of the South Tyrolean German variety and of the entire concept of autonomy, a crucial aspect of the Autonomous Province of Bolzano, has important historical backgrounds that need to be briefly mentioned in order to fully understand the origins and evolution of this linguistic minority.

2.6.1.1. Historical background

After Austria's defeat in World War I, the historical region of Tyrol, which belonged to the Austro-Hungarian Empire, was divided. The Treaty of Saint-Germain-en-Laye established the annexation of the part of Tyrol located South of the main Alpine ridge. This was partly an Italian-speaking region (today's Trentino), while today's South Tyrol was mainly inhabited by German speakers. Even though the following years were characterized by laws aimed at preserving the peculiarities of the German speaking minority, with the advent of fascism, South Tyrol underwent a radical change. This new regime in Italy adopted strong measures to linguistically 'convert' the newly conquered region. The so-called **Italianization** process consisted in a series of measures aimed at denationalizing South Tyrol by forcing cultural and ethnic assimilation and by populating the area with Italian speakers. German surnames were Italianized, the German language was prohibited and Italian was introduced as the only idiom in all public offices and schools. After an agreement between Hitler and Mussolini in 1939, all German speakers were offered the possibility to leave their homeland and emigrate to Germany or to stay in South Tyrol, the region in which their identity was no longer protected. Nevertheless, World War II prevented the complete emigration of German speakers.

After World War II, the first Statute of Autonomy was introduced in 1946, but it was extensively criticized since it failed to create an equal society and to fully grant the right to use the minority language German within the public administration. Finally, the New Statute of Autonomy came into force in 1971, which attributed a separate autonomy to both provinces of South Tyrol and Trentino. Furthermore, the New Statute introduced the right for the citizens to utilize their mother tongue (German or Italian) in their public lives and especially in courthouses and in police or public offices. With the use of German for legal-administrative activities, it became necessary to develop a coherent and correct language with a suitable terminology to fulfil the requirements needed by the citizens and institutions (Chiocchetti et al. 2013).

2.6.1.2. Developing a unique terminology

The process of developing a legal language usually requires many years and the collaboration of different roles (Ralli & Stanizzi 2019). In South Tyrol, this process was not taken into consideration for over 50 years, resulting in an incomplete equalization of German and Italian, due to the lack of an administrative and legal apparatus aiming at utilizing German as an official language as well. For this reason, a modernization of the South Tyrolean bureaucratic system to welcome a new language within it was strictly needed. From a purely linguistic point of view, developing a new South Tyrolean legal language required being able to make it simple (to facilitate understanding) and modern, which meant avoiding calques from Italian and avoiding the use of foreign legal-administrative terms that did not fully reflect the concepts of the Italian legal system and that could therefore create misunderstandings. Every legal system is different and it is made up of specific norms and concepts that are strongly characterized by their uniqueness and by their belonging to legal categories which may be more or less different compared to others. As a consequence, the legal language expresses the concepts of that specific legal system and must conform to the same rules. Even tough Austria, Switzerland, Germany and South Tyrol share the same language, their legal systems are different and terms that share the same surface form may have different meanings. Nevertheless, the process of planning a terminology was not implemented at first. Instead, politics focused on representing both language groups in official bodies and public offices by translating legal and administrative documents, but without a careful planning and collaboration between bodies (Alber & Palermo 2012). This led to high quality translations, which, however, were not terminologically consistent, resulting in different translations for the same terms (Sandrini 1998; Zanon 2001; Chiocchetti et al. 2013). Beside inconsistencies, the lack of planning as well as the lack of expertise in administrative and legal translations resulted in loan words and loan translations, e.g. ‘Quästur’ for Italian ‘questura’ (*police station* in English) or ‘zivile Motorisierung’ for Italian ‘motorizzazione civile’ (*motor vehicle registry* in English). According to Chiocchetti (2021), “[t]erminological awareness was raised by professionals of the law, linguists (not necessarily purists), the media and the education system” (Chiocchetti 2021: 117).

For the purpose of finding a solution to the absence of an appropriate terminological heritage, a Terminology Commission consisting of six members (three German speakers and three Italian speakers) was established in 1988. The objective of this Commission was to “retrieve, check, approve and update the legal, administrative, and technical terminology”

(Chiocchetti 2019: 106). The Commission progressively tackled many administrative and legal sectors, asking for advice to experts in the related fields (e.g. judges and lawyers for the criminal field or employees of the national police for road traffic provisions etc.) (Ralli & Stanizzi 2019: 176). Preliminary terminological research was conducted by the Institute for Applied Linguistics of Eurac Research and the terminological decisions of the Commission were approved by higher authorities before declaring a term as approved. Today, such decisions are visible on the Information System for Legal Terminology *bistro* (n.d.), where standardized terms are marked with the label *Südtirol genormt*, as shown in Fig. 5.

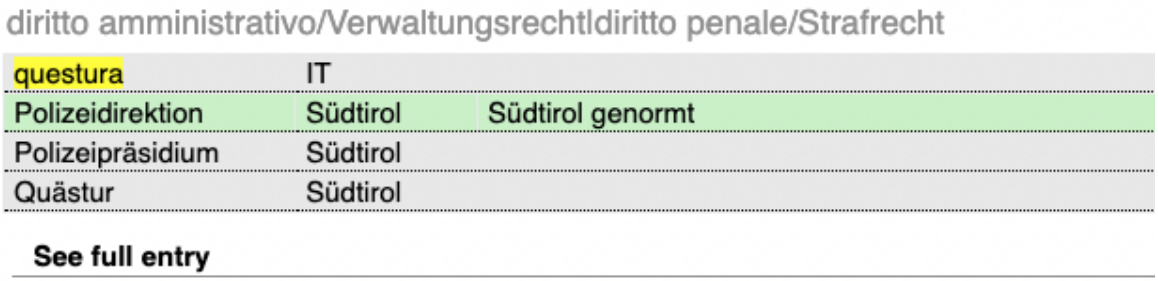

Figure 5: *Südtirol genormt*-label in the bistro entry for the term 'questura'

The work of the Terminology Commission played a crucial role in strengthening the German legal language in South Tyrol and contributed to creating more reliable and consistent translations by regulating more than 20,000 legal terms (Chiocchetti et al. 2013). The terminological development in South Tyrol mainly affected the legal and administrative fields. A particular subdomain of law in which terminology plays an important role is the occupational health and safety domain, which was chosen for the experimental part of this thesis. The next subsection is dedicated to this subfield of law, with particular regard to the South Tyrolean system.

2.6.1.3. Occupational health and safety in South Tyrol

Broadly speaking, occupational health and safety is a field of public health and a subdomain of law that studies hazards in the work environment with the main goal of reducing illnesses and injuries by implementing strategies and regulations and applying them to all workers in a specific country. It also focuses on legal requirements and strictly describes measures to

be adopted in case of noncompliance of regulations. The right to work is enshrined in the Italian Constitution. Since the 1950s, the Italian and European legislation have progressively embraced the concept of occupational health and safety. The focus used to be exclusively on the prevention of accidents and diseases thanks to technological progress (e.g. safer materials and machinery). Today, the attention has shifted toward personal conduct, asking all the people involved to actively participate in training courses in order to improve health and safety levels and reduce occupational diseases and accidents at work (Natullo 2015). This important shift happened thanks to the introduction of several laws adopted over time and collected within legislative decree number 81 of 9 April 2008 (Legislative Decree No. 81/2008), also known as the 'Consolidated Law on Safety at Work' (*Testo Unico sulla Salute e Sicurezza sul Lavoro* in Italian). Information, education and training of employees and in general the human aspect and coordination between the different roles have become essential. Occupational health and safety is a field with high priority in the worker's lives. Nevertheless, even though health and safety levels in Italy are constantly improving, the number of accidents at work, both fatal and non-fatal, is still too high. In 2021, a year that was heavily influenced by the Covid 19 pandemic, a total of 555,236 accidents at work were reported, 1,221 of these were fatal (INAIL 2022). There is still a great deal of work and research to be done to ensure the health and safety of all workers.

In Italy, after a general training course on occupational health and safety, employees are obliged to attend more specific courses according to their tasks and responsibilities. An employee working mostly in an office is likely to attend a specific course on office work to understand and deal with common hazards of the working environment. Additional courses on first aid, firefighting, electrical work, work at height, off-road driving etc. are necessary depending on the employee's duties.

From a terminological point of view, occupational health and safety can be considered a domain with a strong interdisciplinary character. Covering many different topics, this field results in having terminology traits that pertain to multiple subfields. Beside specific concepts and roles pertaining to occupational health and safety only, such as the risk assessment (*Risikobewertung* in German), the personal protective equipment (*persönliche Schutzausrüstung* in German), the coordinator of the health and safety committee (*Leiter des Arbeitsschutzdienstes* in German), other concepts in the field of law, especially the labor law (e.g. the employer, *Arbeitgeber* in German) and the administrative law, gain particular importance (Chiocchetti & Stanizzi 2021: 71). As a subfield of law, occupational health and safety contains a surprisingly large amount of technology-related terms, coming from very

distant domains such as physics (e.g. noise, *Lärm* in German; artificial optical radiation, *künstliche optische Strahlung* in German), chemistry (e.g. explosive mixture, *explosionsfähiges Gemisch* in German), biology (e.g. mutagenic agent, *erbgutverändernder Stoff* in German), mechanical engineering (e.g. emergency stop device, *Not-Halt-Gerät* in German), medicine (e.g. asbestosis, *Asbestose* in German), psychology (e.g. burnout) etc. (Chiocchetti & Stanizzi 2021: 71-72). Nevertheless, being bound to the Italian legal system, creating an accurate terminology in South Tyrolean German has been a complicated task and in many cases terminology gaps due to the absence of a related concept in the German, Austrian or Swiss legal system were present. The translation of the Consolidated Law on Safety at Work was published in 2011, but it has been noted that many terms are not consistently used in all texts on occupational health and safety in South Tyrol. For example, the term 'health and safety representative' was translated with 'Sicherheitssprecher' in the official translation of the Consolidated Law on Safety at Work, but other variants such as 'Arbeitnehmervertreter für Sicherheit' or 'Arbeitnehmervertreter für Arbeitssicherheit' can be found in other sources (INAIL 2018). It has also been observed that acronyms are often not translated and they appear in their Italian form after the German term. In many cases, no actual equivalent for Italian acronyms exists, but also in the presence of a correct acronym in German, the usage of the Italian version prevails. That is the case of the 'Einheitsdokument für die Bewertung der Risiken durch Interferenzen' (*Single Document on the Assessment of Risk from Interference* in English), which is the German equivalent of 'Documento unico di valutazione dei rischi da interferenze'. The corresponding Italian acronym DUVRI was often found in many German texts. Another example is the usage of the Italian acronyms RSPP (short for 'responsabile del servizio di prevenzione e protezione', *coordinator of the health and safety committee* in English) and ASPP (short for 'addetto al serivzio di prevenzione e protezione', *health and safety committee operator* in English) when utilizing the German terms 'Leiter des Arbeitsschutzdienstes' and 'Beauftragter des Arbeitschutzdienstes', which have their corresponding acronyms, i.e., LASD and BASD.

For the purpose of bridging these terminological gaps and making occupational health and safety courses available both in German and Italian, the Institute for Applied Linguistics of Eurac Research translated and revised 31 e-learning modules (that were later distributed to employees for their training) with the main objective of systematizing and spreading a terminology that is clear, coherent and correct, so that every course participant can understand the concepts and apply them in their working lives (Chiocchetti & Stanizzi 2021: 175). These e-learning modules have been translated using CAT tools and the terminology

(a total of more than 1,600 concepts) has been successfully made publicly available on *bistro*. This technology-based approach has contributed to the development and growth of digital translation and terminological resources (more details about the creation of the corpus in Section 4.2). Nevertheless, this approach is not widely adopted in the Autonomous Province of Bolzano and useful data are spread throughout several resources, making South Tyrolean German a low-resource language in the domain of law and administration.

2.6.1.4. South Tyrolean German as a low-resource language

From a generic perspective, a low-resource language can be characterized as such when it is "less studied, resource scarce, less computerized, less privileged" etc. (Magueresse et al. 2020). According to Magueresse et al. (2020), most research carried out on NLP, including machine translation, focuses on a very little percentage of the total number of languages used worldwide, namely on 20 of the 7000 languages spoken around the world. Nevertheless, it is important to state that South Tyrolean German differs from other low-resource languages, since it is considered low-resourced only in specific domains. As mentioned in Section 2.6.1.2, due to its bounding to the Italian legal system, it has developed a different legal and administrative language than Austrian, Swiss and German from Germany and especially its terminology in such domains plays an important role in its definition as low-resourced. German is spoken by roughly 300,000 people in South Tyrol (ASTAT 2021) and the actual translation policies in this region make it difficult to obtain a large quantity of valuable digital resources that could be used for an increasing number of activities in the language industry. Since South Tyrol is an official bilingual province, a number of official legal-administrative documents, such as provincial decrees or acts intended for public use, must be available both in Italian and German. The Office for Language Issues is responsible for the linguistic revision of the provincial decrees and only occasionally deals with the actual translation process (Sandrini 2019: 303). The rest of the documents that must be available in German are translated by individual departments of the administration. Furthermore, such translations are carried out by untrained employees, who in most of the cases are non-professional translators (De Camillis 2021: 204). The use of translation technologies is still underdeveloped, especially in a context where employees find themselves having to consult old translations to find an already translated section. As Sandrini (2019: 111-112) properly points out, the usage of translation technologies is not only beneficial for the quality of the end product and the improved efficiency and productivity of the translators, but also for the

creation of new digital text resources that would highly benefit the growth of a minority language such as South Tyrolean German.

There has been research aiming at augmenting training data for NMT by targeting low-frequency words and generating new, synthetic sentence pairs containing these words (Fadaee et al. 2017). Such approaches may seem feasible in the South Tyrolean German context, but De Camillis (2021) proposed alternative solutions, which would drastically improve the availability of data of this language variety. First of all, there is a strong need to create open-source technologies that are easily accessible for everyone, even if this process requires technical expertise and financial resources. Second, the creation of translation memories is a task with high priority to facilitate the translators' work and to allow the sharing of these memories between offices with the main goal of delivering more reliable translations and possibly create more parallel corpora that can be used to train new MT systems. Many bilingual provincial decrees have already been collected in a platform called *LexBrowser* (n.d.), but it does not allow to consult both Italian and German versions at the same time. A reorganization of the alignment of this platform would allow to better consult the texts and to extract the information in a simplified way. Since the provincial decrees contain the specific legal-administrative South Tyrolean terminology, the amount of texts collected in the *LexBrowser* is the starting point for the creation of adapted MT systems that could later be optimized by inserting further parallel corpora created by the administrative departments (De Camillis 2021: 321). At the moment of writing (2022) there have been two attempts aiming at adapting an existing MT system to the South Tyrolean legal-administrative domain. The first one was carried out by Contarino (2021) with the integration of the *LexBrowser* and the second one is the present master's thesis. Such systems represent a leap to a more modern, uniformed and rapid institutional communication (De Camillis 2021: 321). Finally, De Camillis (2021: 321-322) proposes the development of a linguistic platform in which all provincial decrees, the translation memories of the single departments, the information system *bistro* and possibly adapted MT models can be stored. Such a platform would allow employees to quickly retrieve valuable linguistic data and it would contribute to the continuous development of this minority language.

3. Related work on domain adaptation

MT is a vibrant research field and several scholars have been trying to overcome some of the problems of these types of systems. The goal of domain adaptation is to train or fine-tune an MT model with domain-specific data so that the system is able to learn from these inputs and produce a higher quality target text. Domain adaptation has been carried out on SMT as well and Läubli et al. (2013) demonstrated that even a small number of in-domain data (translation memories) can improve the translation performance if added to a model trained on large out-of-domain datasets. They created two models for two language pairs, namely German-French (DE-FR) and German-Italian (DE-IT), and the specific domain was the automobile industry. They were able to extract 166,957 segments of in-domain data for DE-FR and 112,166 segments for DE-IT. For the out-of-domain data they used freely available corpora and extracted around 8.5 million segments. They then created different models, such as the baseline model with in-domain data only, but also combinations of unweighted models and finally the weighted language and translation model with all the necessary data. They were able to improve the BLEU, TER and METEOR scores in every language pair and a human evaluator chose the results of the weighted language and translation model as the most accurate. Furthermore, adding in-domain data was beneficial for the terminological accuracy as well, since the systems were able to use the right term in the right context. Even if SMT have been superseded by NMT nowadays, the theory behind this approach is still important to understand how domain adaptation in NMT systems works.

Some of the first researchers who were able to adapt an existing NMT system to a new domain were Luong and Manning (2015) and many of the papers listed below took their methods as a reference. At the very beginning of NMT, the authors conducted their experiments on an existing system consisting of 8 individual models trained with 4.5 million parallel sentences. They further trained the system with additional 200,000 sentences of in-domain data in the high-resource language pair German-English to gain 3.8 BLEU points. Furthermore, by ensembling all the models they gained other 2 BLEU points. Another important experiment for domain adaptation in NMT was carried out by Freitag and Al-Onaizan (2016). Their goal was to build an adapted system quickly by adding in-domain data without decreasing the quality on the general domain. Instead of adding in-domain data to the out-of-domain data and then retraining the system from scratch, they re-used the already trained baseline model and continued the training with the small amount of in-domain data. They called this type of training *continue model*. The problem of this model is

that it can cause so-called over-fitting, which means that it overgeneralizes on the training data and thus fails to provide accurate predictions for previously unseen data. To overcome this problem, they ensembled the continue model with the baseline model to get better translations on the new domain without deteriorating the quality of the baseline model. The first experiment was done in the language pair German-English and they used 3.9 million parallel sentences to train the baseline model and added further 194,000 parallel sentences to reach a domain adaptation. They were able to gain 4.4 BLEU points and 3.9 TER points. The second experiment was done in the language pair Chinese-English. The NMT model used was trained with 11.6 million out-of-domain sentences and specialized with 593,000 parallel sentences in order to gain 9.5 BLEU points and 12.2 TER points.

It is evident that the latter three experiments used much more parallel sentences than the amount available for the present work, i.e., roughly 25,000 translation units. Servan et al. (2016) were able to obtain a domain adaptation with far less data and introduced the new concept of *domain specialization*. Their goal was to create domain adaptation in a CAT tool environment, so that adaptation proceeds when the translators are working on their texts in the software and produce new data while translating. Their research is very similar to the one that will be carried out in this master's thesis. They first created a generic model and then specialized it with a corpus from the European Medical Agency. As discussed in Section 2.4.4, they were able to obtain a valuable increase of the BLEU score just by adding a relatively small amount of in-domain data. It is evident that this approach "reaches good performances in a far less time than full-retraining" (Servan et al. 2016) and would be very beneficial for translators using CAT tools and MT systems.

In the papers that were just mentioned, the domain adaptation method of fine-tuning was used. This method could cause an over-fitting problem due to the small size of the training data as reported by Freitag and Al-Onaizan (2016). To overcome this problem, Chu et al. (2017) proposed a domain adaptation method called *mixed fine-tuning*. As always, they first trained an NMT model with out-of-domain parallel sentences, but instead of fine-tuning it on the in-domain corpus, they did it on a parallel corpus, which consisted of a mix of in-domain and out-of-domain corpora. In order to get a comparison with other methods, they performed the experiments with fine-tuning, multi-domain, mixed fine-tuning and combinations of the three. The language pairs used were Chinese-English and Chinese-Japanese to get a mixture of high-resource and low-resource language pairs. The results saw a higher BLEU score in almost every test performed. However, this approach foresees the

development and maintenance of multiple NMT models for different domains and results in increased costs as well as in limits of scalability in real-world scenarios (Farajian et al. 2018).

For this purpose, Farajian et al. (2017) developed a different approach called instance-based adaptation, which aims at maintaining the benefits of corpus-based approaches while "keeping the scalability of one single model" (Farajian et al. 2018: 151). Instead of fine-tuning the whole model, they developed different techniques to perform an instant domain adaptation on sentences similar to the one to be translated retrieved from a parallel corpus (more details about this approach in Section 2.4.4). This method shows valid results in translating domain-specific terminology and is currently used to perform domain adaptation in CAT tools (Bertoldi et al. 2018).

Other research conducted to simulate domain adaptation in a working scenario, e.g. smaller translation companies, freelancers etc., was carried out by Šoštarić et al. (2019) with the goal of comparing domain adaptation results of Google AutoML with an NMT system trained from scratch. The results favored the model trained from scratch, due to more freedom in choosing the best adaptation method and having a better control over the data. Taking into consideration the relatively small amount of in-domain data (42,000 parallel segments), the Google AutoML model was able to increase the BLEU score by 1.6 points but underperformed with the out-of-domain test set, which could be a major problem for companies that want their system to perform well on other domains as well.

Other research on domain adaptation for NMT aims at adapting vocabulary and terminology. Sato et al. (2020) stated that a domain adaptation between distant domains cannot be performed effectively because of the problem of mismatches in vocabulary. This could also be a problem in the specific domain of occupational health and safety since it is a mixture of different domains and words with multiple meanings or whose meaning shifts across domains can be found. To overcome this problem, they proposed the so-called *vocabulary adaptation*, a method that relies on fine-tuning and "adapts embedding layers in a given pre-trained NMT model to the target domain" (Sato et al. 2020: 4269). Building on this new method, researchers were able to increase the BLEU score in the language pairs English-Japanese and German-English by 3.86 and 3.28 points respectively.

4. Methodology

This chapter is dedicated to the methodology adopted to carry out the experimental part of this master's thesis. It starts with the selection of the domain used to carry out the domain adaptation, followed by the experimental settings. Details about the data used to adapt the NMT models, the preprocessing steps adopted to clean the data and the splitting of the dataset into training set and test set are provided. Then follows information on the selection of a pre-trained NMT model, together with the choice of the automatic evaluation metrics and the description of the taxonomy used to manually assess the quality of the adapted model.

4.1. Domain selection

In order to carry out a domain adaptation of an MT model it is vital to select a domain on which the data will focus. Koehn (2020) generally describes a domain "as a collection of text[s] with similar topic, style, level of formality, etc." (Koehn 2020: 239). Occupational health and safety is the domain chosen for the experimental part of this thesis, with particular regard to the Italian and South Tyrolean system.

A discussed in Section 2.6.1.3, occupational health and safety is considered a particularly broad domain, covering many topics and disciplines such as medicine, technology, law, toxicology, ergonomics, materials science etc. Every single field could be split in different domains, each of them containing different in-domain data, but this procedure would go beyond the scope of this thesis, without considering the difficulties of retrieving large quantities of data in a low-resource language pair for adapting an MT model.

Covering many different topics, occupational health and safety results in having peculiar traits such as the presence of very specific terms referring only to this domain, but also terms coming from more distant domains and describing hazardous substances, physical and chemical agents, electrical components etc. Furthermore, being enshrined in the Italian Constitution, occupational health and safety is part of the national legislation. Official provincial documents in the Autonomous Province of Bolzano describing legal requirements and regulations must be available in both official languages, i.e., Italian and German. The Consolidated Law on Safety at Work is not considered a provincial legislation and consequently it is not compulsory to have a German translation. Nevertheless, important documents such as the Consolidated Law on Safety at Work, which have a strong impact on

people's physical safety, are translated for ease of application and transparency. As previously discussed in Section 2.6.1.2, legal terminology must be adapted to a specific legal system. Therefore, many German legal terms pertaining to the occupational health and safety domain are specific to South Tyrol since they express concepts of the Italian legal system. Being a low-resource language variety, an MT model would have several difficulties when confronted with this specific terminology and here is where an adaptive system should succeed.

To the best of my knowledge, this master's thesis is the first attempt to adapt an MT model to the occupational health and safety domain in the language pair Italian and South Tyrolean German. As mentioned in Section 1.2, having an MT model that is specifically trained for that domain would be very beneficial in the Autonomous Province of Bolzano, since new e-learning modules and other texts are continuously written and translated. The peculiarities of this domain, namely a very specific terminology and its interdisciplinary character as a subdomain of law covering many different disciplines, make it the perfect candidate to test the effects of domain adaptation.

4.2. Experimental settings

For the purpose of testing the effects of domain adaptation on different scenarios, it has been decided to create three separate experiments. The **first experiment** aims at translating a text of roughly 100 pages on occupational health and safety made available by the Institute for Applied Linguistics of Eurac Research to test the effects of the domain adaptation on a full document with the presence of out-of-domain sentences as well. Such experiment will give valuable insights on how the system reacts to the presence of out-of-domain segments after having been adapted. The **second experiment** consists in manually extracting domain-specific segments from the text used in the first experiment and testing the effects on shorter and more specific paragraphs to simulate a scenario in which only in-domain segments must be translated. These segments will also serve for the human evaluation. The **third experiment** is carried out to compare the automatic scores with the previous two experiments by randomly extracting a test set from the whole dataset and analyzing whether this common technique leads to higher scores instead of utilizing another unseen text for the translation process. It is common practice to create a test set by extracting it randomly and automatically from the overall dataset. This procedure aims at utilizing only unseen segments for the evaluation and it is considered a *de facto* standard in MT research. The

presence of unseen segments only is not guaranteed in the first two experiments, since there is the possibility that the test sets contain sentences already seen by the model.

4.3. Training and test data

As mentioned in Section 2.6.1.3, training courses on occupational health and safety are mandatory for every employee working in Italy. Today, due to recent changes in Italian legislation, namely the Agreement State-Regions of 26 January 2012 (Agreement State-Regions 26/01/2012), many of these courses, beside those that need practical experience, are held online and employees receive their certificate of completion after having passed a final exam. For this purpose, the Service for Prevention and Safety of the Autonomous Province of Bolzano and the Institute for Applied Linguistics of South Tyrolean research center Eurac Research signed a bilateral agreement with the aim of revising and translating e-learning modules[1] on occupational health and safety in the two official languages of South Tyrol. The goal of this convention is to provide course participants with high-quality content that is stylistically, terminologically and legally coherent. Therefore, a group of terminologists and translators carried out terminology work and efficiently translated and revised 28 e-learning modules, which were later utilized by all public employees in the province. Upon request, such training courses are also made available to private companies.

By translating these e-learning modules in CAT tools, a large translation memory has been created, consisting of roughly 18,000 translation units covering the main topics of occupational health and safety. Furthermore, with South Tyrol being an official bilingual province, all provincial legislative decrees must be available in German as well. The Consolidated Law on Safety at Work is not part of the provincial legislation, but it is considered a central and vital document of the Italian legislation and a corresponding translation in German is strictly necessary. This legislative decree has been aligned with its corresponding equivalent in German, resulting in an additional translation memory containing 6,848 translation units. These two distinct parallel corpora differ in terms of style, genre and level of formality. The e-learning modules aim at educating employees by informing them on details and common procedures regarding occupational health and safety, by providing practical examples and therefore use a register suitable for this purpose, so that every employee is able to understand the content and retrieve all information needed to perform their work safely. Conversely, the Consolidated Law on Safety at Work reflects all

[1] E-learning modules refer to the training courses that can be held online.

characteristics of legal texts: rigid structure (titles, articles, subsections, references to laws etc.), highly syntactic complexity, high register, presence of specific legal terminology etc. The topics of the two documents are alike, they differ however in their purposes and in the resulting textual genre.

Following the work of Contarino (2021) of creating a parallel corpus from the bilingual collection of national and South Tyrolean legislation *LexBrowser*, there was the possibility of utilizing this parallel corpus consisting of 174,468 cleaned sentence pairs as input for the domain adaptation process. The corpus contains high-quality data coming from translations of legal-administrative texts, but the number of data pertaining to the occupational health and safety domain is scarce, many of which are already present in the Consolidated Law on Safety at Work. Therefore, it would not represent a useful resource of in-domain data and it has been decided to not integrate the *LexBrowser* corpus with the already foreseen parallel corpora.

For the evaluation of the adapted NMT model, a text unseen by the model must be utilized. Fortunately, Eurac Research created a new training course destined to internal supervisors. The text is conceptually and terminologically similar to the previous training courses present in the parallel corpora and it has been translated by experienced translators so that there is a reference translation that can be used for the automatic evaluation metrics. We can consider this text as a sort of test set, even if there is no guarantee that all the segments are unique and not previously seen by the model. The second test set used to perform the second experiment consists in very specific in-domain segments on occupational health and safety extracted from the full reference text. The third test set is extracted automatically from the parallel corpora used to adapt the system (details about the dataset splitting in Section 4.3.2) and it is used in the third and last experiment only. The necessary data used to carry out the experimental part of this thesis as well as their size prior to preprocessing are listed in Table 2.

Data used for the experiment	*Name*	*Size*
Parallel corpus 1: e-learning modules on occupational health and safety	SSL	18,583 segments
Parallel corpus 2: Consolidated Law on Safety at Work	TU	6,848 segments
Test set 1: internal training course for supervisors	RefT	1,937 segments
Test set 2: extract of in-domain segments from the internal training course for supervisors	reduced_RefT	79 segments
Test set 3: test set extracted randomly from the parallel corpora	test_set	682 segments

Table 2: data-overview for the experiment

4.3.1. Preprocessing the data

As mentioned in Section 2.4.3.1, text preprocessing consists of three main steps: tokenization, normalization and noise removal. Since the goal of this thesis is not to train an NMT system from scratch and since the parallel corpora used to implement the NMT system during the experimental part of this work are translation memories in a TMX format, the first two steps (i.e., tokenization and normalization) can be discarded. When processing the data, the pre-trained model will directly utilize the .tmx files as input for the domain adaptation, thus avoiding additional preprocessing steps. Even though the majority of the translation memories have been produced and properly curated by professional translators using CAT tools or are the product of a careful alignment of official bilingual texts, it has been proven that noise can drastically reduce the quality of MT, especially NMT (Belinkov & Bisk 2017; Khayrallah & Koehn 2018).

Contarino (2021) developed a toolkit[2] written in Python to remove noise from the *LexBrowser* parallel corpus created for his research and since a part of the data used to adapt the models comes from similar sources and has the same language pair, it has been decided to use it for the present research endeavor as well. The toolkit consists of a class *ParallelCorpus* with 10 additional functions. Some of them perform cleaning operations on the segments without actually deleting the translation units, whereas others are designed to

[2] https://github.com/antcont/LEXB/blob/master/mt/tmx_cleaner.py

completely remove noisy translation units from the parallel corpora. The functions used to remove noise are the following:

- **remove_untranslated**: it removes translation units where the source segment and the target segment are identical or almost identical. As previously discussed in Section 2.4.3.1, this is considered the noise type with the most catastrophic impact on the overall quality, since it learns copying behaviors and apply it to other segments.
- **remove_useless**: it removes useless translation units containing only elements such as 'Art. 1', '(1)', '1.', '1bis'. It is particularly useful for the parallel corpus of the Consolidated Law on Safety at Work.
- **non_alphabetical_ratio_filter**: it removes translation units having a large quantity of punctuation, digits and whitespaces and filters out segments with a non-alphabetical to alphabetical characters ratio bigger than 0.8 as well.
- **noise_cleaning**: a method for carrying out common noise removal such as deleting quotes, brackets, bullet points, numbers at the beginning or at the end of a segment, as well as other punctuation and digits.
- **dehyphenation**: it analyzes words with the objective of finding out whether a hyphenated word is more frequent as hyphenated or merged. If it finds a hyphenated word that should be merged, it converts the word to its merged equivalent.
- **remove_blank_units**: it removes translation units where the source segment does not have a corresponding translation, hence the target segment is empty.
- **remove_whitespaces**: it removes leading and trailing whitespaces in the translation units.
- **language_filter**: it detects and removes translation units consisting in a language pair that is not Italian-German.
- **length_ratio_filter**: it removes translation units which have a length ratio higher than a given threshold, considering them misaligned sentences.
- **filter_per_token:** it removes very short and very long segments according to a predefined number of tokens.

4.3.2. Dataset splitting

For the third experiment of the present work (see Section 4.2 for a description of the experiments), a test set is extracted from the overall dataset to test the effects of the domain adaptation using a more common technique for evaluating the quality of the NMT models, instead of utilizing an unseen text as a reference translation. This procedure is commonly adopted in machine and deep learning experiments, where the overall datasets used to create ML and DL models are usually divided into training set, validation set and test set. The training set corresponds to the data used to train the model. The validation set is the set of data used to validate the results that the model returns during training, useful for fine-tuning the model's hyperparameters and determining the state of its quality (Géron 2019). Finally, the test set is the set of data unseen by the model that serves for the final evaluation (Zafar et al. 2018: 202). The latter is only used once the model is completely trained. It is common practice to utilize 80% of the data for training, 10% for validation and 10% for testing, but taking into consideration the overall amount of data available for the present work, it has been decided to use only 5% for the test set, since a larger quantity would reduce the training data and might affect the quality of the model. Since a pre-trained model is used and no training from scratch is performed, the validation set does not need to be extracted and 95% of the data can be used for training/adaptation.

The extraction must be random and for this purpose, a Python script created by Contarino (2021) and publicly available on GitHub[3] is utilized. The script builds on the method of Pinnis (2018) and aims at performing an advanced deduplication operation to avoid having duplicates and near-duplicates in the test set. Pinnis (2018) decided to completely remove near-duplicates from the dataset, but the following script was designed for smaller datasets and removing such translation units from a corpus containing roughly 23,000 segments could influence the overall performance. For this purpose, near-duplicates are temporarily stored in a blacklist to avoid inserting them in the test set and reinserted in the training set once the test set has been successfully extracted. The script also performs several normalization operations such as lowercasing words, replacing digits with a placeholder, removing whitespaces and punctuation. The Python script is used on both parallel corpora, i.e., SSL and TU. Considering the size difference between the corpora, 3,5% is extracted from the SSL corpus and 1,5% from the TU corpus. The sentences that

[3] https://github.com/antcont/LEXB/blob/master/mt/dataset_deduplication%2Bsplitting.py

form the test set are translation units with 10-20 words length starting with uppercase and ending with a period, colon or semicolon.

Since the script only supports tab separated txt files, the corpora must be converted into the desired format and re-converted into TMX for the domain adaptation process. For this purpose, a powerful TM tool called Heartsome TMX Editor[4] is used. The software allows to perform various operations on TMX files, including converting the parallel corpora into a txt format and re-converting the txt files into TMX.

4.4. Domain adaptation

Having selected a domain, retrieved the necessary data, preprocessed them to remove noisy elements that could deteriorate the quality of the model and split the dataset into training set and test set, the domain adaptation process can start. Before carrying out a domain adaptation, it is necessary to select a pre-trained model and the next two subsections are dedicated to this process.

4.4.1. Selecting a pre-trained model

Domain adaptation can be performed in different ways. Many attempts aimed at training a system from scratch with a large amount of data before proceeding with an additional training with in-domain data or running a few iterations on in-domain data only to reach a domain adaptation. These methods have shown valid results in both SMT and NMT (Läubli et al. 2013; Luong & Manning 2015; Servan et al. 2016; Freitag & Al-Onaizan 2016). The downside of these approaches is that a great deal of computation power as well as time is required to train large MT models from scratch or to fine-tune the system to the chosen domain. Creating a machine/deep learning model also requires technical expertise and even though many toolkits that support developers have been created, many of them do not allow domain adaptation, especially if needed in a short amount of time.

OpenNMT (Klein et al. 2017) is an open-source toolkit for neural machine translation which allows users to create customized NMT models by providing implementations in two popular deep learning frameworks, i.e., *PyTorch* and *Tensorflow*. OpenNMT offers a variety of features, including pre-trained models. It includes state-of-the-art translation models based on Transformer and Bidirectional Long Short-Term Memory (BiLSTM) architectures

[4] https://github.com/heartsome/tmxeditor8

only in the language pairs English-German, German-English and Catalan to other languages. Creating an NMT model with low-resource language pairs would require a training from scratch. The advantage of toolkits such as OpenNMT is the possibility of creating new systems with in-domain data only, if in possession of a large quantity of parallel sentences. If this approach seems feasible for larger companies that possess computational and financial resources as well as professional competences in the field of natural language processing and MT, it may represent a barrier for smaller translation and language companies and freelance translators, who would benefit from the availability of pre-trained models that can be customized to a particular domain with less data and without particular programming skills (Šoštarić et al. 2020). For this purpose, systems such as *Google AutoML* or *Microsoft Translator HUB* have been developed. Google AutoML, for example, runs on the Google Cloud and enables customers to create their own customized machine learning models by providing parallel data in a TMX or TSV format. Google AutoML provides pre-trained models coming from the generic Translation API model and further accepts user's data to deliver a domain adaptation. It offers basic preprocessing steps such as noise removal, i.e., discarding duplicate sentences, very long sentences, sentences with identical source and target, sentences with a wrong language pair etc. and dataset splitting. If the user does not specify which data to use for the training set, the validation set and the test set, Google AutoML will automatically use 80% of the data for training, 10% for validation and 10% for testing. The automatic metric used for evaluating the model is BLEU.

Nevertheless, even though such customizable systems offer valuable features that highly facilitate the creation and maintenance of ML models, they only support limited language combinations, the majority of them containing English, thus removing the possibility of utilizing pre-trained models if other language pairs want to be used. For the present research endeavor, the new adaptive NMT system called *ModernMT*, which offers the language pair Italian-German and further improves the performances shown by similar systems such as Google AutoML is used to perform a domain adaptation.

4.4.2. ModernMT

ModernMT (Bertoldi et al. 2018) is a cloud-based adaptive NMT system which accepts multiple language pairs and allows users to quickly customize NMT models to a particular domain by adding TMs and consequently adapting terminology, style etc. Furthermore, thanks to its integration as plug-in in several CAT tools, it allows a real-time learning from

human corrections. The most valuable distinction from previously mentioned systems such as Google AutoML is the new technology on which ModernMT is based. ModernMT uses a so-called *document-level translation*, which takes into consideration the content of the whole document to be translated instead of focusing on single sentences thanks to a so-called *context-vector*. The context-vector is a representation of the context of a document with reference to the stored memories. Each memory has a corresponding weight, which can have a maximum value of 1. The closer the weight is to 1, the more the memory content will influence the translation result. As previously mentioned in Section 2.4.4, ModernMT is based on the *instance-based adaptation* proposed by Farajian et al. (2017) and uses a Fairseq Transformer architecture (Ott et al. 2019). As demonstrated by Farajian et al. (2018), the instance-based adaptation shows valid results in translating domain-specific terminology, with automatic scores comparable to the ones obtained by corpus-based approaches, but with significantly less time involved in the adaptation process, since instance-based adaptation fine-tunes the model on the fly.

ModernMT is open source[5], but if users do not want to train their systems from scratch, a license is needed to download the commercial product and make use of pre-trained models for several language pairs. For the experimental part of this thesis, the commercial product is used. The cleaned parallel data are uploaded on the ModernMT plug-in on SDL Trados Studio, together with the test set that needs to be translated. Since the instance-based approach performs the adaptation on the fly, the upload only requires a short amount of time. After the upload, the NMT model is ready to be deployed. ModernMT automatically translates the whole text, which can further be exported in a txt format and imported in the script that will perform the automatic evaluation. The translation carried out by the baseline model is obtained by translating the test set in the generic API system available online (ModernMT n.d.). This process is repeated for all three experiments described in Section 4.2.

[5] https://github.com/modernmt/modernmt

4.5. Quality evaluation

Once ModernMT has translated the test sets with its adapted model and with the baseline model, the evaluation process consisting of an automatic and a human evaluation can start. The next two subsections are dedicated to the choice of the automatic metrics and the creation of a taxonomy to perform a more detailed human evaluation.

4.5.1. Automatic evaluation – BLEU and METEOR

As mentioned in Section 2.5.1, several automatic metrics have been developed over the years to assess MT quality in a fast and inexpensive way. After performing a domain adaptation with ModernMT, the results of the adapted and the baseline model are evaluated using two automatic metrics: BLEU and METEOR.

Despite the already mentioned limitations of BLEU, e.g. penalizing translations with different lengths and the usage of synonyms, it is still considered the *de facto* standard metric and offers a great comparison with other research in the field of MT. Nevertheless, Post (2018) demonstrated that BLEU scores do not always allow comparability due to variations in the internal parameters of the metric and differences in the preprocessing steps of the reference and candidate translations. To offer comparability across studies, he developed a modification called *SacreBLEU*. This metric, which still builds on the original paper of Papineni et al. (2002), is used together with METEOR to assess the quality of the adapted and the baseline NMT model. METEOR scores are reported because they offer more precision thanks to the combination of unigram matching, unigram precision and unigram recall and support the usage of synonyms and morphological variants (Banarjee & Lavie 2005). Both automatic metrics are computed with Python using the *dataset* library from the open-source machine learning toolkit *Hugging Face* and the *NLTK* library for natural language processing for the tokenization of the texts.

4.5.2. Human evaluation – taxonomy

The BLEU and METEOR scores are particularly useful to quickly assess MT quality. Nevertheless, such metrics are very literal and slight changes might affect the overall score. Translating in a non-literal way, merging and/or splitting sentences to better express the meaning of the source text in the target language can deeply affect the scores of the automatic metrics. Furthermore, they fail at analyzing complex linguistic phenomena and do not

provide information about the nature of errors. In order to carry out a complete evaluation, a human assessment is strictly necessary to discover where the model has improved and what type of errors it delivered. Translation quality of the text produced by the adapted model is analyzed by annotating translation errors. The taxonomy used for the human evaluation is based on the work of Tezcan et al. (2017) and divides error categories into two main groups: accuracy and fluency. As mentioned in Section 2.5.2, accuracy is analyzed by looking at the relationship between source and target text, whereas fluency focuses on the target text and is concerned with its construction and with the correctness of the target language. By analyzing an extract of the adapted translation consisting of 79 segments, several types of errors are annotated. The types of errors for both accuracy and fluency are listed in Table 3.

Accuracy Errors		**Fluency Errors**	
Addition		Grammar	Multi-Word Syntax Word Form Word Order Extra Words Missing Words Other
Omission		Lexicon	Non-existing or Foreign Word Lexical Choice
Untranslated		Orthography	Spelling Capitalization Punctuation Other
Do-Not-Translate (DNT)		Multiple Errors	
Mistranslation	Multi-Word-Expressions Part-Of-Speech (POS) Word Sense Disambiguation (WSD) Partial Semantically Unrelated Gender Other	Coherence	Co-reference Consistency
Mechanical	Capitalization Punctuation Other		

Bilingual Terminology			
Source Error			
Other			

Table 3: type of errors to annotate during the evaluation

In more details, *Addition* errors refer to the target text not being present in the source (not applicable to punctuation), whereas *Omission* errors refer to the cases in which the source content cannot be found in the target. *Untranslated* errors can be annotated when the source text is not translated but is copied to the target segment instead. *Do-Not-Translate (DNS)* errors refer to the source content which is unnecessarily translated into the target language (e.g. proper nouns). *Mistranslation* errors are further divided into multiple categories (see Table 3 for an exhaustive list) and happen when the source content is translated incorrectly (not applicable to bilingual terminology). When accuracy errors not related to the transfer of meaning such as punctuation and capitalization are identified, they can be included in the category of *Mechanical* errors. An important error category is *Bilingual Terminology*, which refers to the incorrect translation of predefined bilingual terminology for South Tyrol. This means that the target term may be a correct translation of the source term, but it is not the preferred one in the Autonomous Province of Bolzano. The correctness of such terms is validated through internal terminology databases or through the publicly available information system for legal terminology *bistro*. Finally, *source errors* refer to errors that are present in the source segment and are not originated by the MT system.

Errors that can be annotated under fluency are related to the construction of the target language. *Grammar* errors refer to the misuse of grammatical rules of the German language and can be further divided into multiple categories (see Table 3 for an exhaustive list). *Lexicon* errors are annotated when the German lexicon is used incorrectly. Subcategories are *Non-existing or Foreign Word*, i.e., a word that does not belong to the German lexicon is used, *Lexical Choice* i.e., the word used is part of the German lexicon, but in order to generate a correct sentence, another word should be utilized. *Orthography* errors refer to the German writing conventions and can be further divided into *Spelling*, *Capitalization* and *Punctuation* errors. *Multiple Errors* are a combination of errors and their presence could create difficulties when trying to annotate single fluency errors separately. An example could be "a word order error combined with wrong word forms and wrong lexical choices" (Tezcan

et al. 2017: 226). The last category under fluency is *Coherence*. It is used to annotate errors referring to coherence at a text level (e.g. inconsistency). To identify this type of errors it is necessary take into account larger parts of a text or the entire text. For a detailed description of all error subcategories see Tezcan et al. (2017).

5. Results

This chapter is dedicated to a detailed analysis of the results of the present work. It starts with the results of the preprocessing steps adopted to clean the data used for the experiments, in detail, the number of segments deleted with each function is given. Results of the dataset splitting of the overall dataset follow. Thereafter, a comparative analysis of the overall quality improvement between the baseline model and the adapted model in all three experiments in terms of BLEU and METEOR scores is given, followed by a detailed human analysis of 79 segments translated by the adapted engine. The test set is analyzed using the taxonomy described in Section 4.5.2.

5.1. Results of the preprocessing steps

Following the methods described in Section 4.3.1 of preprocessing the data used for the experiments, this section is dedicated to the results obtained by applying the toolkit for corpus cleaning and filtering on both SSL and TU corpora.

As already mentioned, the toolkit used to clean the data has 10 functions that clean and filter a parallel corpus as input. After a first analysis of the effects of the toolkit and taking into consideration the overall size of the corpora, the style of the translation units and their purpose, it has been decided to discard the *filter_per_token* function, which removes too short and too long segments. After a first cleaning test of the SSL corpus, the number of segments removed by this function was too high, considering the overall amount of data. It removed 6,101 translation units, equal to roughly 34% of the total amount of data. According to Khayrallah and Koehn (2018), who proved that segments with less than 5 tokens could lower the BLEU score, the function was set to filter out segments containing less than 6 tokens and/or those containing more than 80 tokens. This threshold could easily be modified, but it has been decided to not utilize the *filter_per_token* function for several reasons: the SSL corpus is composed by training courses, whose main purpose is to educate employees on matters pertaining to occupational health and safety. Therefore, many lists with bullet points were produced to facilitate understanding. The function used for eliminating short and long segments would remove all these translation units and its implementation could lead to even worse consequences by drastically reducing the amount of data and by discarding valuable segments containing the domain-specific terminology. It has also been decided to not use it on the TU corpus for the same reasons.

The two parallel corpora have been cleaned with the other functions mentioned in Section 4.3.1 and a total of 1,217 translation units were removed from the SSL corpus, whereas 469 translation units were discarded in the TU corpus.

Initially, the translation units in the SSL corpus where 18,583, but due to formatting issues, the file could not be processed correctly by the Python toolkit. After a new export on SDL Trados Studio without formatting, around 500 units were discarded, resulting in a total amount of 18,009 segments. Most units discarded during the cleaning process were untranslated sentences (target same as source), sentences with a large non-alphabetical to alphabetical characters ratio and sentences with a highly different length ratio.

In the TU corpus, the most frequent noisy elements were again untranslated sentences (the majority of them are name of substances, bacteria etc. that have been left in their official Latin name) and sentences with a highly different length ratio between source and target segments. Detailed results of the noise-cleaning phase are illustrated in Table 4. The total amount of cleaned translation units used for the domain adaption is 23,171.

Type of noise	***Segments of the SSL corpus***	***Segments of the TU corpus***	
Raw corpus	18,009	6,848	
Untranslated (identical source-target)	282	291	
Highly similar source-target	92	34	
Missing translation	0	0	
Wrong language	6	1	
Non alphabetical character ratio	155	34	
Useless	0	0	
Sentence length ratio difference	682	109	
Cleaned corpus	**16,792**	**6,379**	**Tot: 23,171**

Table 4: noise-cleaning operations on the SSL and TU corpora

5.2. Results of the dataset splitting

As described in Section 4.3.2, the entire dataset used for the experimental part of this thesis, i.e., both parallel corpora, has been split into training and test set to carry out the third and last experiment. Splitting datasets in training and test set is a widely adopted method in many ML and DL tasks and it has been decided to test the effects of this technique on both parallel

corpora to analyze whether this common procedure leads to a more substantial improvement of automatic scores. The goal of this procedure is to randomly extract and remove 5% of the overall parallel segments from the dataset to form a test set that will serve for the final evaluation of the model. As mentioned before, taking into consideration the different size of the parallel corpora, it has been decided to extract 3,5% from the SSL corpus and 1,5% from the TU corpus.

Since the script used to split the data was designed to perform an advanced deduplication to avoid having duplicates and near-duplicates in the test set, the corpora underwent a simple deduplication at the beginning, which resulted in a small volume reduction. After that, near-duplicates have temporarily been stored in a blacklist to avoid inserting them in the test set and then re-inserted in the training set. The result is a new training corpus and a new text (the test set) with its corresponding reference translation that can be translated by the baseline and the adapted model.

A total number of 682 translation units were extracted, 587 from the SSL corpus and 95 from the TU corpus. Details about the dataset splitting procedures are listed in Table 5.

Dataset splitting operations	***Segments of the SSL corpus***	***Segments of the TU corpus***
Segments of total dataset before deduplication	16,792	6,379
Segments of total dataset after simple deduplication	16,779	5,841
Number of almost duplicates	255	86
Near duplicates identified in the dataset	560	228
Total segments valid for the test set[6]	2,444	4,39
Test set size	95	587
Training test size	16,193	5,747

Table 5: results of the dataset splitting operations

[6] Refers to the identified segments between 10 and 20 tokens, starting with uppercase and ending with a dot, a colon or a semicolon.

5.3. Results of the automatic evaluation

In order to obtain the first evidence of the quality improvement of the adapted system, all test sets have been evaluated automatically with BLEU and METOR and the scores of the adapted system have been compared with those of the baseline ModernMT system. To obtain a translation performed by the baseline system, in all three experiments carried out to test the effects of domain adaptation, the generic and free translation tool available online has been used. Translations obtained by the adapted system have been carried out on SDL Trados Studio with the ModernMT plug-in trained with both parallel corpora.

Results are shown in Table 6. The highest BLEU and METEOR scores are marked in bold. As can be noted, the highest scores all belong to the adapted models and in all three experiments, the adapted model was able to improve the translation quality. The highest score in terms of BLEU was reported in the first experiment, i.e., the translation of the full e-learning module for supervisors, with 35.72 points and an improvement of more than 2 BLEU points over the baseline model, whereas the highest METEOR score was surprisingly reported in the second experiment, i.e., the translation of 79 in-domain segments extracted from the longer e-learning module for supervisors, with 0.421 points. However, the greatest improvement of the adapted model was reported in the third experiment, i.e., the translation of the test set extracted randomly from the parallel corpora, with an increase of BLEU and METEOR scores by 6.4 and 3.5 points respectively.

	First experiment		**Second experiment**		**Third experiment**	
	Baseline	Adapted	Baseline	Adapted	Baseline	Adapted
BLEU	33.11	**35.72**	28.98	32.52	21.30	27.75
METEOR	0.308	0.322	0.408	**0.421**	0.329	0.364

Table 6: overall quality improvement in terms of BLEU and METEOR scores

The common technique of evaluating the quality through a test set extracted randomly from the entire dataset led to a considerable increase of the BLEU and METEOR scores, which tells us that the type of test set plays an important role while computing these automatic metrics. With only 22,000 translation units used to fine-tune the engine, the model was able

to significantly improve both BLEU and METEOR scores. The overall scores, especially the one of the baseline model (21.30), are considerably low for NMT standards. The reason for such low scores can be the fact that the test set contained roughly 100 segments of the TU corpus, which consists of the translation of the Consolidated Law on Safety at Work. The language of this document is highly specialized and reflects all characteristics of legal language. State-of-the-art NMT still fails at providing users with high-quality legal translations (Heiss & Soffritti 2018; Wiesmann 2019; De Camillis 2021), but this experiment is the proof that fine-tuning a system to a specific domain can improve the quality even with a small number of in-domain data.

In the first and second experiment, the improvement of the automatic scores is not as significant as the one obtained in the third experiment, but the individual scores are certainly higher. The adapted model was able to gain 2.6 BLEU points and 1.4 METEOR points in the first experiment and 3.5 BLEU points and 1.3 METEOR points in the second experiment. Overall, the BLEU and METEOR scores are not particularly high for NMT standards, but the improvement is in line with other experiments conducted on domain adaptation (Freitag & Al-Onaizan 2016; Luong & Manning 2015; Servan et al. 2016) and even surpasses research carried out with far more in-domain data in terms of score improvement. Another important result obtained by the adapted model is the fact that ModernMT did not experience a so-called overfitting when translating the full text on occupational health and safety. The latter contained out-of-domain segments as well, but the model was still able to improve the BLEU score by 2.6 points, thus avoiding a catastrophic forgetting with a consequent quality reduction when facing segments that are not present in the adaptation data, a problem that is very common when fine-tuning an NMT system.

Even though automatic metrics are a valuable resource to quickly assess MT results, the present quantitative analysis through BLEU and METEOR confirms that it is not reasonable to completely rely on automatic metrics to assess MT quality. As demonstrated, differences in the type of test set, style of the translation etc. deeply influence the resulting scores. Furthermore, they do not provide us with results regarding important translation features, such as grammar, terminology, accuracy, etc. For this purpose, the 79 in-domain segments extracted from the full text used in the first experiment have been manually analyzed using the taxonomy described in Section 4.5.2 to better understand where the model improved and what kind of errors it delivered.

5.4. Results of the human evaluation

After the quantitative analysis through BLEU and METEOR, the 79 segments used for the second experiment have been manually evaluated by annotating errors in a platform called *WebAnno* developed by the Technical University of Darmstadt for this exact purpose. Before diving into the analysis of several examples, a statistic of the most occurring errors is given. Following the explanation of accuracy and fluency errors presented in Section 4.5.2, the results are divided into these two categories for both the quantitative and qualitative analysis. Fig. 6 shows the amount of accuracy errors annotated during the human evaluation as well as the type of Mistranslation errors produced by the machine.

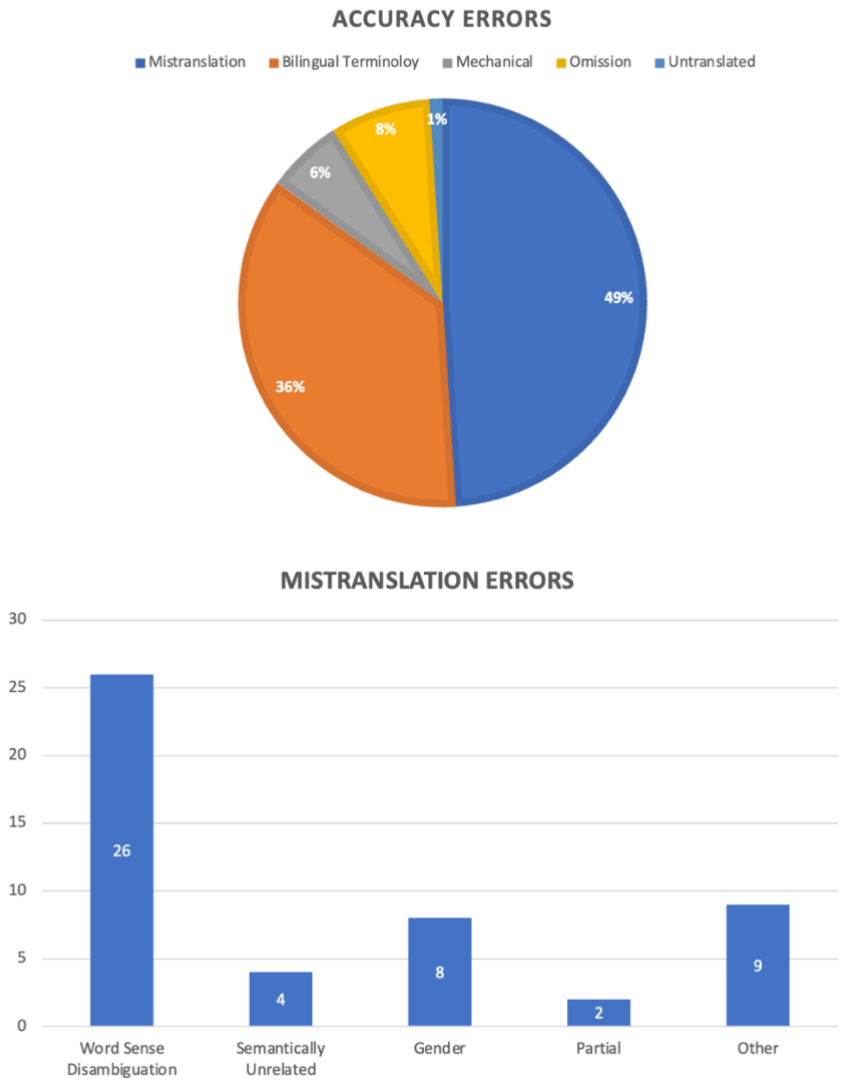

Figure 6: accuracy errors and mistranslations annotated during the evaluation

The adapted model, during the translation of the test set used for the second experiment, produced a total of 100 accuracy errors, 49 of which belong to the category Mistranslation (49%), 36 to Bilingual Terminology (36%), 6 to Mechanical (6%), 8 to Omission (8%) and 1 to Untranslated (1%). Since Mistranslation errors are further divided into several sub-categories and are considered a particularly important category, their occurrence has been analyzed as well. Of the 49 errors annotated, 26 belong to the category Word Sense Disambiguation, 4 to Semantically Unrelated, 8 to Gender, 2 to Partial and 9 to Other. The number of fluency errors is lower compared to accuracy errors and during the evaluation, a total of 40 errors was annotated. As shown in Fig. 7, 16 errors belong to the category Coherence (40%), 14 to Grammar (35%) and 10 to Lexicon (25%).

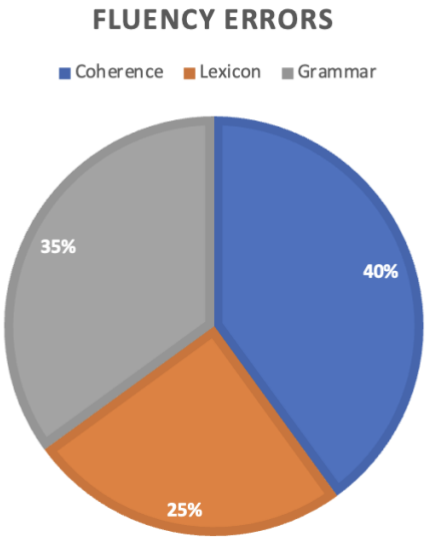

Figure 7: fluency errors annotated during the evaluation

The number of errors produced by the model, especially mistranslations, is still high, even after a domain adaptation. In fact, the goal of domain adaptation is not to completely eliminate every fluency or accuracy error, instead it focuses on reducing these errors by adapting and improving terminology, style, etc. By comparing the translations produced by the adapted and the baseline model, it becomes clear that the domain adaptation process was proven to be effective, especially when faced with domain-specific terminology, which is a particularly important feature when translating texts pertaining to the occupational health and safety domain and the legal domain in general. Table 7 demonstrates the success of the

domain adaptation in translating very specific terms of occupational health and safety correctly.

Source text	Le misure di prevenzione e protezione definite dal datore di lavoro in collaborazione con il Servizio di prevenzione e protezione devono essere applicate dall'intera organizzazione.
Adapted translation	Die vom Arbeitgeber in Zusammenarbeit mit dem Arbeitsschutzdienst festgelegten Arbeitsschutzmaßnahmen müssen von der gesamten Organisation umgesetzt werden.
Baseline translation	Die vom Arbeitgeber in Zusammenarbeit mit dem Präventions- und Schutzdienst definierten Präventions- und Schutzmassnahmen müssen von der gesamten Organisation umgesetzt werden.

Table 7: translation example 1

As can be noted, there is a clear difference between the two translations, not in terms of style or structure, but in terms of terminology. Domain-specific terms, such as 'Arbeitsschutzdienst' (*health and safety committee* in English) and 'Arbeitsschutzmaßnahmen' (*preventive and protective measures* in English), are essential concepts of the occupational health and safety domain and frequently occur in both parallel corpora used to adapt the system as well as in many other texts pertaining to this domain. Comparing the translations produced by the adapted and the baseline models offers a great overview of the use of the domain-specific terminology. The terms translated by the baseline model are not completely wrong and a reader would still be able to understand the meaning, but they are not the right terms used in South Tyrol and therefore need to be annotated as Bilingual Terminology errors. These sentences perfectly exemplify the potential of domain adaptation within companies or for low-resource languages, since vanilla MT represented by the baseline model here would not succeed in translating this type of terminology correctly.

Another example of excellent rendering of bilingual terminology by the adapted MT model is shown in Table 8.

Source text	È quindi fondamentale che tutti i quasi-infortuni siano segnalati al Servizio di prevenzione e protezione, così può analizzarli e apportare i necessari cambiamenti affinché simili eventi (ed eventi più gravi) non si ripetano in futuro.
Adapted translation	Es ist daher wichtig, dass alle Beinaheunfälle dem Arbeitsschutzdienst gemeldet werden, damit dieser sie analysieren und die notwendigen Änderungen vornehmen kann, damit sich solche Ereignisse (und schwerwiegendere Ereignisse) in Zukunft nicht wiederholen.
Baseline translation	Es ist daher wichtig, dass alle Quasi-Unfälle dem Präventions- und Schutzdienst gemeldet werden, damit er sie analysieren und die notwendigen Änderungen vornehmen kann, damit sich solche Ereignisse (und schwerwiegendere Ereignisse) in Zukunft nicht wiederholen.

Table 8: translation example 2

'Beinahe-Unfall' (*near miss* in English) is another important term in the occupational health and safety domain, which refers to any event related to work which may have caused an accident, a harm to health or even death but, only by mere chance did not cause it. The adapted model recognized the term from the adaptation data and translated it correctly. Conversely, the baseline model did not have any adaptation data available but relied instead on its general training data and produced an incorrect translation. The latter not only translated it incorrectly, but misinterpreted the Italian word 'quasi', resulting in multiple errors within the same word, i.e., Bilingual Terminology under accuracy and Lexical Choice under fluency. Furthermore, it translated once again 'Servizio di prevenzione e protezione' incorrectly, error committed throughout the entire translation. The example sentences of Table 8 also show that the domain adaptation did not provide an improvement of the linguistic style, since it is clear the adapted and the baseline translations only differ in terms of terminological accuracy.

As mentioned in Section 4.4.2, ModernMT uses a so-called instance-based adaptation that consists in adapting the model's hyperparameters only when the algorithm recognizes a similar sentence present in the training/adaptation data. That results in an adaptation of single sentences and not of the entire model. Nevertheless, this approach can be considered particularly useful when translating very repetitive texts with several definitions, as could be legal documents or, for instance, training courses on occupational

health and safety. The following sentences in Table 9 provide a good example of the success of this method.

Source text	Definizione: persona in possesso delle capacità e dei necessari requisiti professionali designata dal datore o dalla datrice di lavoro, a cui risponde, per coordinare il Servizio di prevenzione e protezione dai rischi.
Adapted translation	Definition: Person, die über die beruflichen Fähigkeiten und Qualifikationen verfügt, um den Arbeitsschutzdienst zu koordinieren, und die dem Arbeitgeber untersteht.
Baseline translation	Definition: Person, die über die erforderlichen beruflichen Fähigkeiten und Anforderungen verfügt und vom Arbeitgeber oder der Arbeitgeberin benannt wurde, um den Risikopräventions- und -schutzdienst zu koordinieren.
Reference translation	Definition: Person, die über die beruflichen Fähigkeiten und Qualifikationen verfügt, um den Arbeitsschutzdienst zu koordinieren, und die dem Arbeitgeber bzw. der Arbeitgeberin untersteht.
Reference in the parallel corpus	Persona in possesso delle capacità e dei necessari requisiti professionali designata dal datore di lavoro, a cui risponde, per coordinare il servizio di prevenzione e protezione dai rischi.

Table 9: translation example 3

Noticing the difference not only in the terminology, but also in the construction of the phrase, it becomes clear that the algorithm saw a very similar sentence in the adaptation data and copied it to produce a very good translation, almost identical to the reference translation. In fact, the definition of coordinator of the health and safety committee can be found in the Consolidated Law on Safety at Work as well as in the e-learning modules translated by Eurac Research, which were both used as parallel corpora to adapt the model. The reference in the corpus is shown in Table 9 and it demonstrates that the system recognized the almost perfect match and decided to use it for the translation. Nevertheless, the translation produced by the adapted model presents a Gender and an Omission error since it failed to translate the feminine noun of 'datrice di lavoro' (*employer* in English). This example shows the tendency of ModernMT of ignoring important features of the source text after finding a good match in the adaptation data. In fact, all definitions present in the Consolidated Law on Safety at

Work and in many e-learning modules too, have not been gendered, whereas it has been decided to use both the masculine and feminine form or gender-neutral nouns and pronouns in the reference text and in its corresponding reference translation.

Gender bias in MT is a common problem (Prates et al. 2019; Stanovski et al. 2019) and there have been studies aiming at reducing gender bias through domain adaptation (Saunders & Byrne 2020). Both translations produced by the baseline and the adapted models often failed in rendering gendered nouns, pronouns and adjectives correctly. As seen in Table 9, one main reason is the fact that the language present in the adaptation data was not gendered enough and the machine was not able to adapt these traits. Table 10 shows another example of gender errors in the adapted translation.

Source text	MEDICO O MEDICA COMPETENTE Definizione: professionista in medicina in possesso di uno dei titoli e dei requisiti formativi e professionali necessari, che collabora con il datore o la datrice di lavoro ai fini della valutazione dei rischi ed è nominato/a dal datore o dalla datrice di lavoro per effettuare la sorveglianza sanitaria e altri compiti relativi alla salute del personale
Adapted translation	BETRIEBSARZT ODER BETRIEBSARZT Definition: Mediziner, der über einen der erforderlichen Titel, beruflichen Befähigungen und Ausbildungen verfügt, mit dem Arbeitgeber oder der Arbeitgeberin zusammenarbeitet, um die Risikobewertung durchzuführen, und der vom Arbeitgeber oder der Arbeitgeberin mit der Durchführung der Gesundheitsüberwachung und anderen mit der Gesundheit des Personals zusammenhängenden Aufgaben betraut ist.

Table 10: translation example 4

In the source text, beside explicitly mentioning both gender forms of the word ‘medico competente’ (*appointed doctor* in English), it has been decided to opt for a gender-neutral form in the definition. This detail was not taken into consideration in the adapted translation and the machine decided to use the masculine noun ‘Mediziner’ as well as masculine articles in the relative clauses. Already in the title, even though the noun was explicitly used in both forms in the source text, the MT system was not able to recognize this feature and translated it twice in the masculine form, resulting not only in a Gender error under accuracy, but also in a Grammar error (Extra Words) under fluency. However, in the definition under the title

it used both masculine and feminine forms when translating ‘il datore o la datrice di lavoro’, which shows us the inconsistency in producing reliable translations.

Another problem observed is the inability for the model to look at a longer textual sequence and taking into consideration previous sentences instead of focusing on single segments. This resulted in coherence errors, since many segments are strongly correlated to previous sentences or words, as could be the introduction of a list. Table 11 shows how the sentences (in detail, the verb forms) have been translated without taking into consideration how the list was introduced. These mismatches between entities across sentences do not make the sentences themselves grammatically incorrect, but surely incorrect on a paragraph or text level.

Source text	È tuo compito: - verificare che le sostanze in uso siano le meno pericolose a disposizione per quel determinato utilizzo; quando possibile, sostituire le sostanze con altre meno pericolose; segnalare la possibilità di sostituzione a dirigente, datore di lavoro e al Servizio di prevenzione e protezione - segnalare a dirigente, datore di lavoro e al Servizio di prevenzione e protezione qualsiasi situazione di potenziale rischio legata all’uso di sostanze chimiche, biologiche, cancerogene e mutagene - conoscere le sostanze chimiche, biologiche, cancerogene e mutagene
Adapted translation	Es ist Ihre Aufgabe: - Überprüfen Sie, ob die verwendeten Substanzen für diese Verwendung am wenigsten gefährlich sind; wenn möglich, ersetzen Sie die Substanzen durch andere, die weniger gefährlich sind; melden Sie die Möglichkeit des Austauschs an Führungskräfte, Arbeitgeber und den Arbeitsschutzdienst - Unterrichtung des Managements, des Arbeitgebers und des Arbeitsschutzdienstes über alle potenziellen Risikosituationen im Zusammenhang mit der Verwendung von chemischen, biologischen, krebserzeugenden und erbgutverändernden Stoffen - die chemischen, biologischen, krebserregenden und erbgutverändernden Stoffe kennen

Table 11: translation example 5

By looking at the single sentences of the list, no coherence errors are to be seen, nevertheless, by considering how the list was introduced, it becomes clear that the construction of the following sentences is wrong and it is incorrect to start a sentence with 'Überprüfen Sie' after having introduced it with 'Es ist Ihre Aufgabe'. The right translation would be 'zu überprüfen'. Therefore, it is necessary to annotate them as Coherence (Co-Reference) errors under fluency, since it is not necessary to look at the source text to classify them as errors. Of course, many other errors are present in the following sentences, such as the transformation of the verb 'segnalare' into the noun 'Unterrichtung', which would fall under the category of Mistranslation (Part of Speech) or the translation of 'È tuo compito' with 'Es ist Ihre Aufgabe' instead of 'Es ist deine Aufgabe'. These types of errors, i.e., using the polite form in German even though it was not used in the source text, frequently occurred throughout the entire translation of the test set, as illustrated in Table 12.

Source text	Ti devi quindi sempre accertare (come anche i dirigenti e le dirigenti) che il Servizio di prevenzione e protezione sia stato PREVENTIVAMENTE informato dell'introduzione di nuovi macchinari, attrezzature o strumenti di lavoro. In questo modo il Servizio può valutare i rischi e definire le misure di prevenzione e protezione necessarie PRIMA CHE AVVENGA LA MODIFICA.
Adapted translation	Sie müssen daher immer sicherstellen, dass der Arbeitsschutzdienst VORAB über die Einführung neuer Arbeitsmaschinen, Arbeitsmittel oder Arbeitsmittel informiert wurde. Auf diese Weise kann der Dienst die Risiken bewerten und die erforderlichen Arbeitsschutzmaßnahmen definieren, BEVOR DIE ÄNDERUNG ERFOLGT.

Table 12: translation example 6

The source text does not use the polite form, which was adopted in the translation. These types of errors are not particularly critical, but surely represent a mistranslation that could lead to an ambiguity and are therefore annotated as Mistranslation (Word Sense Disambiguation). Furthermore, if not used consistently throughout the text they could cause inconsistencies on a fluency level. Nevertheless, the segments of Table 12 present other errors that the domain adaptation was not able to solve, partially because of ambiguities in the source text, such as the presence of both 'attrezzature' and 'strumenti di lavoro', which are usually translated with 'Arbeitsmittel' but in this case caused a fluency error, i.e., Extra Words, but also due to a lack of context-understanding. The word 'Servizio' (notice the

capitalized ‘S’) refers to ‘Servizio di prevenzione e protezione’, which was always translated correctly with ‘Arbeitsschutzdienst’. This is a typical situation of translating from a high-context language (Italian) into a low-context language (German). In this case, the Italian source text uses an ellipsis or a hyperonym of the term, but rendering it into German requires precision, otherwise the translation results in misunderstandings. Humans have astounding capabilities of understanding context from very few examples, a task that a machine fails to obtain. The NMT system was not able to understand the context and translated the term only with ‘Dienst’ (*service* in English).

Other errors that MT systems often produced are omissions. For instance, in the example of Table 12, the part of the source text included in the parenthesis has not been translated. Together with omissions, typical errors of MT systems are untranslated words, sentences etc., which refer to the source text been copied to the target text. This phenomenon is illustrated in Table 13 and is very common when translating acronyms. In this case, the acronym for ‘datore di lavoro’ has been left in the source language, that is ‘DL’. A professional human translator would have probably decided to not translate it or to use the correct acronym in German, i.e., ‘AG’. Nevertheless, as discussed in Section 2.6.1.3, this error could be motivated by the fact that many German texts on occupational health and safety in South Tyrol do not use a corresponding German acronym for a specific term, instead, the Italian acronym is left untouched.

Source text	DATORE O DATRICE DI LAVORO (DL)
Adapted translation	ARBEITGEBER ODER ARBEITGEBERIN (DL)

Table 13: translation example 7

As mentioned in the first translation example of Table 7, the domain adaptation process led to a considerable increase of the terminological accuracy. Nevertheless, the adapted model did not succeed in translating all domain-specific terminology correctly and still produced 36 Bilingual Terminology errors. Compared to the baseline model, which failed at rendering the majority of the terms correctly, the adapted model underwent a considerable improvement, which nonetheless was not consistent throughout the entire translation. As can be noted in the example sentences of Table 14, three domain-specific terms have not been

translated correctly, especially if used in a text pertaining to the South Tyrolean language variety. All three mistranslations fall under the category of Bilingual Terminology errors. According to the information system for legal terminology *bistro* (n.d.), 'videoterminalista' should have been translated into 'Bildschirmarbeiter' (*DSE worker* in English), the correct translation of 'Gesundheitsprotokoll' would have been 'ärztliches Untersuchungsprotokoll' and 'medico competente' should have been rendered with 'Betriebsarzt'. The latter was previously translated correctly, but in this context the machine did not recognize the term and decided to translate it with 'zuständiger Arzt' instead.

Source text	Tutto il personale è considerato videoterminalista e sottoposto a visite mediche periodiche, come da protocollo sanitario specifico elaborato dal medico competente in base ai rischi per la salute rilevati.
Adapted translation	Alle Mitarbeiter gelten als Videoterminalisten und werden regelmäßig medizinisch untersucht, gemäß dem spezifischen Gesundheitsprotokoll, das vom zuständigen Arzt auf der Grundlage der festgestellten Gesundheitsrisiken erstellt wird.

Table 14: translation example 8

One main issue that MT is not able to completely solve is dealing with words with multiple meanings. Table 15 shows this peculiarity when the system was asked to translate the verb 'diluire'. ModernMT translated it literally with 'verdünnen' (*dilute* in English), which certainly causes a misunderstanding and would therefore need to be annotated as a Mistranslation (Word Sense Disambiguation).

Source text	Chi lavora è più esposto alle cause della patologia rispetto al resto della popolazione. Tra le possibili cause diluite nel tempo si conosce, ad esempio, l'esposizione a: sostanze chimiche, agenti biologici, rumore, vibrazioni, radiazioni, videoterminali.
Adapted translation	Wer arbeitet, ist den Ursachen der Krankheit stärker ausgesetzt als der Rest der Bevölkerung. Mögliche Ursachen, die im Laufe der Zeit verdünnt werden können, sind z. B. die Exposition gegenüber: Chemikalien, biologischen Stoffen, Lärm, Vibrationen, Strahlung und Videoterminals.

Table 15: translation example 9

Beside the Word Sense Disambiguation error, the sentence also presents a Bilingual Terminology error in the translation of ‘videoterminalista’ with ‘Videoterminalisten’, as previously seen in the example in Table 14.

The previous 9 examples served as a demonstration of the effects of the domain adaptation on an in-domain test set, the full translations are provided in Appendix 1. As demonstrated, the model was able to quickly adapt important translation features such as terminology and occasionally the style. Important terms of the occupational health and safety domain have been learned by the adapted model and translated correctly more often than by the baseline model, albeit not consistently. Nevertheless, as shown in Figures 5 and 6, the amount of errors, especially Mistranslations is still high even after adapting the system and inconsistencies in the use of bilingual terminology throughout the translation are visible. The results of the human evaluation are in line with those of the automatic evaluation in that the improvement of the adapted model is noticeable and taking into consideration the relatively small amount of data used for the adaptation, it is appropriate to underline the great potential of domain adaptation within the field of MT. Having more data available would play a crucial role in possibly obtaining an even more evident quality improvement.

6. Discussion

Based on the results presented in Sections 5.3 and 5.4, it becomes clear that the methodology adopted to address the research question and to perform the domain adaptation was proven to be effective. The adapted model was able to improve the automatic scores in all three experiments. In the first two experiments, an improvement could be observed, but it is not particularly high, whereas in the third experiment the adapted model was able to improve the BLEU and METEOR scores by 6.4 and 3.5 points respectively, but without reaching high scores. In the first experiment, ModernMT delivered the highest BLEU score overall with 35.72 points. In general, the improvement of the automatic scores obtained in the third experiment is in line with other research on domain adaptation of NMT models. For instance, the research conducted by Servan et al. (2016) and presented in Section 3, shows very similar results. On a test set extracted randomly from the corpus, they were able to obtain an improvement of 7.53 BLEU points, but on the model trained with 50,000 parallel segments, i.e., double the amount of data available for the present master's thesis. A further interesting comparison are the results obtained by Šoštarić et al. (2019), who tested the effects of domain adaptation on the commercial product *Google AutoML*. They used 42,000 parallel segments coming from the medical domain in the language pair English-Croatian and further utilized 12,000 medical terms for the domain adaptation, but obtained worse results compared to the present research, namely an improvement of only 1.6 BLEU points.

Another interesting result worth of discussion is the difference of BLEU and METEOR scores when computing them on different types of test sets. The results presented in Section 5.3 confirm that the type of test set highly influences the resulting scores. These observations put into question the reliability of the majority of the results in terms of automatic scores if applied to real-world scenarios. In fact, the goal of machine translation is to deliver high-quality translations in a real-world scenario, namely by translating a long and structured text, as was carried out in the first experiment. Clearly, the quality improvement of the adapted model in terms of BLEU and METEOR scores compared to the baseline model in the first experiment is far less impressive than the one obtained in the third experiment, which was the translation of an unseen test set extracted randomly from the parallel corpora. The latter test set consisted in 682 random segments with a length of 10-20 tokens. Even though such test sets are widely adopted when evaluating MT models, they cannot be considered actual texts, since they lack context, structure etc. Nevertheless, these test sets are the *de facto* standard in MT research and even though the results of the automatic evaluation in the first

two experiments are particularly adapted if applied to a real-world scenario, they do not provide truthful scores for machine or deep learning experiments, since they contained duplicate sentence pairs that distort the calculation of the automated metrics. Nevertheless, the test sets used in the first two experiments offer a further valuable feature: professional translators often find themselves having to re-translate similar sentences. When using translation technologies, this is done automatically by CAT tools, but these experiments proved that ModernMT is able to obtain the same results. As shown in Table 9 of Section 5.4, ModernMT found a good match in its adaptation data and applied it to the translation. This so-called instance-based adaptation could represent an enormous advantage when asking the machine to translate very repetitive texts, as could be legal-administrative texts, if trained with the correct data that contain similar sentences and could highly support post-editing processes.

One question that arises is the ability for non-experts in computer science to replicate the results obtained in the present research. The present work is an interesting approach for a master's thesis conducted in a translation studies department, since computer science skills were required to fine-tune the NMT models and to make use of the data. Specifically, procedures to clean the data and to split them into training set and test set have been adopted. Such processes require a basic understanding of computer science and programming, but many open-source toolkits are available on the web, so that it is not necessary to develop new scripts from scratch. Furthermore, even though it has been proven that noisy data highly influence the quality of NMT (Belinkov & Bisk 2017; Khayrallah & Koehn 2018), if the parallel corpora come from translations produced by means of CAT tools and have been constantly curated, they could be inserted into the system without further cleaning them. This is also the goal of ModernMT, i.e., facilitating the creation of adapted NMT models without deep technical competences, which are not particularly common within the translators' community. Moreover, the dataset splitting operation is performed to evaluate the quality of the model by means of automatic metrics. In a real-world scenario, such process is not strictly required and the model could be instantly deployed to translate new texts.

Regarding the human evaluation, it offered valuable insights regarding translation quality that could not be delivered by automatic metrics. The results presented in Section 5.4 and the analysis of 79 segments confirm that the quality of the adapted model improved, especially in important traits such as bilingual terminology and occasionally style. Many typical terms of the occupational health and safety domain as well as official bilingual terms

of South Tyrol have been successfully learned by the adapted model and adopted in the translation. As expected, the baseline model did not succeed in this task and was not able to translate the very specific terms correctly. According to previous research (Heiss and Soffritti 2018; Wiesmann 2019; De Camillis 2021), translating South Tyrolean legal-administrative terminology correctly is one of the main issues of MT systems in the context of the Autonomous Province of Bolzano. Nevertheless, a quick analysis of Fig. 6 in Section 5.4 confirms that the terminology improvements are not particularly striking and the presence of 36 Bilingual Terminology errors is the proof that more efforts need to be put into ensuring a more precise translation of such terms. Terminological inconsistencies have also been observed during the human evaluation (see Table 14 of Section 5.4) and terms that have been translated correctly in one segment have been rendered incorrectly in another segment. Of course, the small amount of in-domain data surely played a crucial role in this aspect as well, since it is likely that the model did not have a sufficient amount of data from which to learn the terminology and other important translation features.

In fact, the main limitation of this work is undoubtedly the reduced amount of data. Many published experiments on domain adaptation were carried out with a substantial larger amount of data (Luong & Manning 2015; Freitag & Al-Onaizan 2016), namely 200,000 or even 590,000 parallel segments. The corpus used to adapt the models contains roughly 23,000 cleaned parallel segments, making its size very limited for MT training purposes. Moreover, even if the data contained in the parallel corpora are translations carried out by professionals or are official bilingual documents, they could be further improved. For instance, the decisions of not gendering the majority of names and adjectives in the Consolidated Law on Safety at Work and in many e-learning modules present in the parallel corpora were certainly reflected in the resulting translations. The usage of a more gender specific language would certainly enhance the translation performances.

7. Conclusions

The starting point of the present research was that machine translation, although having reached surprising quality levels, still fails when confronted with very specific domains and with low-resource language varieties. To address the research question, the extent to which the quality of an existing NMT system improves when performing a domain adaptation in the language pair Italian-South Tyrolean German in the very specific domain of occupational health and safety was analyzed. The motivation of this research was to create an adapted model able to assist the continuous need for new translations in the legal-administrative domain in South Tyrol and to highlight the importance of South Tyrolean legal-administrative terminology in the context of MT as well as to estimate the size of data required for domain adaptation. To answer the research question, two parallel corpora containing a total of roughly 25,000 in-domain translation units were made available by South Tyrolean research center Eurac Research. The corpus was then preprocessed and cleaned to avoid inserting noisy data into the system. 23,000 cleaned translation units have been inserted into the NMT system ModernMT, which performed the domain adaptation on three different scenarios. For the first two experiments, an unseen reference text created by Eurac Research was used, whereas for the third experiment a test set extracted randomly from the parallel corpora was used. All test sets have been translated by both the baseline and the adapted ModernMT models and the translations produced by the machine have been evaluated automatically by means of automatic metrics and manually through a taxonomy aimed at identifying major error categories.

The results have shown promising improvements by having incremented the BLEU and METEOR scores in all three experiments and by having reached an important increase in terminological accuracy, which nevertheless could be further improved. Many bilingual terms have been translated correctly by the adapted model, but inconsistencies are noticeable and the adaptation is not visible throughout the entire translation.

The present work contributed to the already vibrant research field of MT, with a special focus on the South Tyrolean context. This research estimated the amount of data needed for domain adaptation in a low-resource language setting and demonstrated that even a small amount of in-domain data can contribute to increasing the quality of NMT without the need to completely retrain a system from scratch, which would require computational resources, time and a large amount of data. It also demonstrated that a domain adaptation setting could be the perfect solution for small translation companies and freelance translators with limited

resources. Their translation memories in a TMX file represent the only resource they would need and without particular competences in the field of computer science they could create their own customized NMT models. In detail, the present work contributed to creating an adapted model able to translate legal-administrative documents on occupational health and safety in the languages Italian and South Tyrolean German. Even though its quality can be further improved by future research, it represents a good starting point to apply MT in the translation workflow.

Future research may overcome some of the limitations introduced in Section 6, especially through the initiatives proposed by De Camillis (2021) and described in Section 2.6.1.4. Thanks to a systematic translation process in CAT tools, the amount of high-quality data can be expanded and such resources could build the foundation to creating a system adapted with far more data, which would highly benefit its improvement. Furthermore, testing new kinds of domain adaptation, e.g. integrating terminology constraints as presented in Section 3 or synthetically augmenting the training data for low-resource languages, could offer valuable insights. An interesting approach, which would, however, require highly technical competences in the field of natural language processing, would be creating a custom NMT system from scratch trained with data pertaining to the South Tyrolean legal-administrative domain and comparing the results with the already existing adapted ModernMT systems, i.e., the research carried out by Contarino (2021) and the present master's thesis. However, this last proposal would need an extraordinarily high amount of data currently not available. Nevertheless, by applying a more standardized translation policy together with the use of translation technologies, the amount of high-quality data would increase exponentially, resulting in more research opportunities. Finally, in order to apply MT in the translation workflow in South Tyrol there is the need to further investigate its feasibility and its success in terms of improved productivity, ideally in experiments with professional translators.

8. Bibliography

Aggarwal, Charu C. (2018) *Neural Networks and Deep Learning: A Textbook*. New York: Springer.

Agreement State-Regions 26/01/2012 – Accordo Stato Regioni 26/01/2012 "Diagramma di Flusso sui Percorsi e le Attività di Formazione e Sicurezza".

Alansary, Sameh (2014). Interlingua-based Machine Translation Systems: UNL versus other interlinguas. *The Egyptian Journal of Language Engineering*, *1*(1), 42-54.

Alber, Elisabeth & Palermo, Francesco (2012). Creating, Studying and Experimenting with Bilingual Law in South Tyrol: Lost in Interpretation? In: Xabier, A. (ed.) *Bilingual Higher Education in the Legal Context. Group Rights, State Policies and Globalisation*. Leiden: Nijhoff, 287–309.

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.: National Academy of Sciences, National Research Council, 1966. (Publication 1416.).

Bahdanau, Dzmitry; Cho, Kyunghyun & Bengio, Yoshua (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR abs/1409.0473*.

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, Michigan: Association for Computational Linguistics, 65–72.

Barbu, Eduard; Parra Escartín, Carla; Bentivogli, Luisa; Negri, Matteo; Turchi, Marco; Orasan, Constatin & Federico, Marcello (2016). The first automatic translation memory cleaning shared task. *Machine Translation*, 30(3), 145-166.

Belinkov, Yonatan & Bisk, Yonatan (2017). Synthetic and Natural Noise Both Break Neural Machine Translation. *CoRR abs/1711.02173*.

Bennett, Winfield S. & Slocum, Jonathan (1985). The LRC Machine Translation System. *Computational Linguistics*, 11(2-3), 111-121.

Bertoldi, Nicola; Caroselli, Davide & Federico, Marcello (2018). The ModernMT Project. In *Proceedings of The 21st Annual Conference of the European Association for Machine Translation.* Universitat d'Alacant, Alacant, Spain, 345.

Brown, Peter F.; Della Pietra, Stephen A.; Della Pietra, Vincent J. & Mercer, Robert L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational linguistics*, *19*(2), 263-311.

Callison-Burch, Chris; Fordyce, Cameron; Koehn, Philipp; Monz, Christof & Schroeder, Josh (2007). (Meta-) Evaluation of Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, 136-158.

Castilho, Sheila; Doherty, Stephen; Gaspari, Federico $ Moorkens, Joss (2018). Approaches to Human and Machine Translation Quality Assessment. In: Moorkens. J.; Castilho, S.; Gaspari, F. & Doherty S. (eds) *Translation Quality Assessment. From Principle to Practice*. Cham: Springer International Publishing, 9–38.

Chiocchetti, Elena (2019). Implementing Language Rights in South Tyrol: Developing Legal Terminology for the Public Administration and the Judiciary. In: Horvath, I. (ed.) *Proceedings of the 20th International Translation Studies Conference Organised by the Department of Translation and Interpreting of ELTE University (22−23 March 2018, Budapest, Hungary)*. Budapest: ELTE Eötvös Kiadól, 103–117.

Chiocchetti, Elena; Kranebitter, Klara; Ralli, Natascia & Stanizzi, Isabella (2013). Deutsch ist nicht gleich Deutsch. Eine terminologische Analyse zu den Besonderheiten der deutschen Rechtssprache in Südtirol. In: Brambilla, M; Gerdes, J & Messina, C. (eds.) *Diatopische Variation in der deutschen Rechtssprache*. Frank & Timme, 253-285.

Chiocchetti, Elena & Stanizzi, Isabella (2021). La terminologia della salute e sicurezza sul lavoro in tedesco per l'Alto Adige. In: Grimaldi, C. & Zanola, M. T. (eds.) *Terminologie e vocabolari. Lessici specialistici e tesauri, glossari e dizionari*. Firenze University Press, 71-82.

Cho, Kyunghyun; Van Merriënboer, Bart; Gulcehre, Caglar; Bahdanau, Dzmitry; Bougares, Fethi; Schwenk, Holger & Bengio, Yoshua (2014). Learning Phrase Representations Using RNN Encoder- Decoder for Statistical Machine Translation. *CoRR abs/1406.1078*.

Chu, Chenhui & Wang, Rui (2018). A Survey of Domain Adaptation for Neural Machine Translation. *CoRR abs/1806.00258*.

Clyne, Michael (2012). Pluricentric Languages – Introduction. In: Clyne, M. (ed.). *Pluricentric Languages: Differing Norms in Different Nations*. Boston: De Gruyter Mouton, 1-10.

Contarino, Antonio Giovanni (2021). *Neural machine translation adaptation and automatic terminology evaluation: a case study on Italian and South Tyrolean German legal texts*. Master's Thesis, Alma Mater Studiorum Università di Bologna.

De Camillis, Flavia (2021). *La traduzione non professionale nelle istituzioni pubbliche dei territori di lingua minoritaria: Il caso di studio dell'amministrazione della Provincia autonoma di Bolzano*. Doctoral thesis, Alma Mater Studiorum Università di Bologna.

Dinu, Georgiana; Mathur, Prashant; Federico, Marcello & Al-Onaizan, Yaser (2019). Training Neural Machine Translation To Apply Terminology Constraints. *CoRR abs/1906.01105*.

Fadaee, Marzieh; Bisazza, Arianna & Monz, Christof (2017). Data Augmentation for Low-Resource Neural Machine Translation. *CoRR abs/1705.00440*.

Farajian, M. Amin; Bertoldi, Nicola; Negri, Matteo; Turchi, Marco & Federico, Marcello (2018). Evaluation of Terminology Translation in Instance-Based Neural MT Adaptation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*. Alicante: European Association for Machine Translation, 149–158.

Farajian, M. Amin; Turchi, Marco; Negri, Matteo & Federico, Marcello (2017). Multi-Domain Neural Machine Translation through Unsupervised Adaptation. In *Proceedings of the Second Conference on Machine Translation*. Copenhagen: Association for Computational Linguistics, 127– 137.

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

Freitag, Markus & Al-Onaizan, Yaser (2016). Fast Domain Adaptation for Neural Machine Translation. *CoRR abs/1612.06897.*

Géron, Aurélien (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow.* 2nd Edition. Sebastopol, CA: O'Reilly Media.

Gupta, Rohit; Lambert, Patrik; Patel, Raj & Tinsley, John (2019). Improving Robustness in Real-World Neural Machine Translation Engines. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*. Dublin, Ireland: European Association for Machine Translation, 142–148.

Hangya, Viktor & Fraser, Alexander (2018). An Unsupervised System for Parallel Corpus Filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Brussels, Belgium*:* Association for Computational Linguistics, 882–887.

Hasler, Eva; De Gispert, Adrià; Iglesias, Gonzalo & Byrne, Bill (2018). Neural Machine Translation Decoding with Terminology Constraints. *CoRR abs/1805.03750.*

Hearne, Mary & Way, Andy (2011). Statistical Machine Translation: A Guide for Linguists and Translators. *Language and Linguistics Compass,* 5(5), 205-226.

Heiss, Christine & Soffritti, Marcello (2018). DeepL Traduttore e didattica della traduzione dall'italiano in tedesco – Alcune valutazioni preliminari. *inTRAlinea Special Issue: Translation And Interpreting for Language Learners (TAIL)*, 241-258.

Kalchbrenner, Nal & Blunsom, Phil (2013). Recurrent Continuous Translation Models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1700-170.

Khayrallah, Huda & Koehn, Philipp (2018). On the Impact of Various Types of Noise on Neural Machine Translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. Melbourne, Australia: Association for Computational Linguistics, 74–83.

Klein, Guillaume; Kim, Yoon; Deng, Yuntian; Senellart, Jean; Rush, Alexander M. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*. Vancouver, Canada: Association for Computational Linguistics, 67-72.

Kloss, Heinz (1978). *Die Entwicklung neuer germanischer Kultursprachen seit 1800*. Düsseldorf: Pädagogischer Verlag Schwann.

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

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

Koehn, Philipp; Hoang, Hieu; Birch, Alexandra; Callison-Burch, Chris; Federico, Marcello; Bertoldi, Nicola; Cowan, Brooke; Shen, Wade; Moran, Christine; Zens, Richard; Dyer, Chris; Bojar, Ondřej; Constantin, Alexandra & Herbst, Eva (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, 177-180.

Koehn, Philipp & Knowles, Rebecca (2017). Six Challenges for Neural Machine Translation. *CoRR abs/1706.03872*.

Koehn, Philipp; Och, Franz J. & Marcu, Daniel (2003). *Statistical Phrase-Based Translation*. University of Southern California Marina Del Rey Information Sciences Inst.

Kurfali, Murathan & Östling, Robert (2019). Noisy Parallel Corpus Filtering through Projected Word Embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*. Florence, Italy: Association for Computational Linguistics, 277–281.

Läubli, Samuel; Fishel, Mark; Volk, Martin & Weibel, Manuela (2013). Combining Statistical Machine Translation and Translation Memories with Domain Adaptation. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, 331-341.

Legislative Decree No. 81/2008 – Decreto legislativo del 9 aprile 2008, n. 81 "Attuazione dell'articolo 1 della legge 3 agosto 2007, n. 123, in materia di tutela della salute e della sicurezza nei luoghi di lavoro".

Lommel, Arle; Uszkoreit, Hans & Burchardt, Aljoscha (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista Tradumàtica: tecnologies de la traducció* 12: 455–463.

Luong, Minh-Thang & Manning, Christopher D. (2015). Stanford Neural Machine Translation Systems for Spoken Language Domains. In *Proceedings of the international workshop on spoken language translation*.

Macken, Lieve; Prou, Daniel & Tezcan, Arda (2020). Quantifying the Effect of Machine Translation in a High-Quality Human Translation Production Process. In *Informatics*, vol. 7, Multidisciplinary Digital Publishing Institute, 1-19.

Magueresse, Alexandre; Carles, Vincent & Heetderks, Evan (2020). Low-resource Languages: A Review of Past Work and Future Challenges. *CoRR abs/2006.07264*.

Marcu, Daniel & Wong, William (2002). A Phrase-Based, Joint Probability Model for Statistical Machine Translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, 133-139.

Marie, Benjamin; Fujita, Atsushi & Rubino, Raphael (2021). Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers. *CoRR abs/2106.15195*.

Melby, Alan K. & Wright, Sue Ellen (2015). Translation Memory. In: Chan, Sin-Wai (ed.). *The Routledge Encyclopedia of Translation Technology*. London: Routledge, 662–677.

Mikolov, Tomas; Chen, Kai; Corrado, Greg & Dean, Jeffrey (2013). Efficient Estimation of Word Representations in Vector Space. *CoRR abs/1301.3781*.

Natullo, Gaetano (2015). *Salute e sicurezza sul lavoro*. UTET Giuridica.

Nießen, Sonja; Och, Franz Josef; Leusch, Gregor & Ney, Hermann (2000). An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. Athens, Greece: European Language Resources Association (ELRA).

Ott, Myle; Edunov, Sergey; Baevski, Alexei; Fan, Angela; Gross, Sam; Ng, Nathan; Grangier, David & Auli, Michael (2019). FAIRSEQ: A Fast, Extensible Toolkit for Sequence Modeling. *CoRR abs/1904.01038*.

Palmer, D. David (2010). Text Preprocessing. In: Indurkhya, N. & Damerau, F. J. (eds) *Handbook of Natural Language Processing*. Boca Raton: CRC Press, 9-30.

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 of the Association for Computational Linguistics*, 311-318.

Pinnis, Mārcis (2018). Tilde's Parallel Corpus Filtering Methods for WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, Belgium, Brussels: Association for Computational Linguistics, 939–945.

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

Popović, Maja (2018). Error Classification and Analysis for Machine Translation Quality Assessment. In: Moorkens. J.; Castilho, S.; Gaspari, F. & Doherty S. (eds) *Translation Quality Assessment. From Principle to Practice.* Cham: Springer International Publishing, 129-158.

Post, Matt (2018). A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium: Association for Computational Linguistics, 186–191.

Prates, Marcelo O.; Avelar, Pedro H. & Lamb, Luís C. (2019). Assessing Gender Bias in Machine Translation: A Case Study with Google Translate. *Neural Computing and Applications*, 32(10), 6363-6381.

Ralli, Natascia & Stanizzi Isabella (2018). Il linguaggio giuridico tedesco in Alto Adige. Evoluzione delle politiche terminologiche. *AIDA Informazioni, 36,* 169-189.

Rarrick, Spencer; Quirk, Chris & Lewis, Will (2011). MT Detection in Web-Scraped Parallel Corpora. In *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*. Xiamen, China: International Association for Machine Translation, 422-430.

Rikters, Matiss (2018). Impact of Corpora Quality on Neural Machine Translation. In: Muischnek, K & Müürisep, K. (eds.) *Human Language Technologies - the Baltic Perspective - Proceedings of the Eighth International Conference Baltic HLT 2018*. Tartu, Estonia: IOS Press, 126–133.

Rivera-Trigueros, Irene (2021). Machine translation systems and quality assessment: a systematic review. In *Language Resources and Evaluation*, 1-27.

Rosenblatt, Frank (1957). *The Perceptron, a Perceiving and Recognizing Automaton. Technical report.* Buffalo, NY: Cornell Aeronautical Laboratory.

Rumelhart, David E.; Hinton, Geoffrey E. &. Williams, Ronald J. (1986). Learning internal representations by error propagation. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 1*, 318–362.

Sandrini, Peter (1998). Italienisches Recht in deutscher Sprache. Terminologische Überlegungen. In: Cordin, P.; Iliescu, M. & Siller-Runggaldier, H. (eds.) *Italienisch und Deutsch im Kontakt und Vergleich: Akten des 7. Treffens der Italienischen und österreichischen Linguisten*. Trento: Università di Trento, 399–417.

Sandrini, Peter (2012). Tecnologia FLOSS per la traduzione – Disponibilita, applicazione e problematiche. *Intralinea. Special Issue: Specialised Translation II*.

Sandrini, Peter (2019). *Translationspolitik für Regional- oder Minderheitensprachen – Unter besonderer Berücksichtigung einer Strategie der Offenheit*. Berlin: Frank & Timme.

Saunders, Danielle (2021). Domain Adaptation and Multi-Domain Adaptation for Neural Machine Translation: A Survey. *CoRR abs/2104.06951*.

Saunders, Danielle & Byrne, Bill (2020). Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. *CoRR abs/2004.04498*.

Scott, Bernard E. (1989). The Logos System. In *Proceedings of Machine Translation Summit II*, 137-142.

Servan, Christophe; Crego, Josep & Senellart, Jean (2016). Domain specialization: a post-training domain adaptation for Neural Machine Translation. *CoRR abs/1612.06141*.

Shiwen, Yu & Xiaojing, Bai (2014). Rule-Based Machine Translation. In: Chan, Sin-Wai (ed.) *The Routledge Encyclopedia of Translation Technology*. London: Routledge, 224-238.

Snover, Matthew; Dorr, Bonnie; Schwartz, Rich; 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: Technical Papers*. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, 223–231.

Šoštarić, Margita; Pavlović, Nataša & Boltužić, Filip (2019). Domain Adaptation for Machine Translation Involving a Low-Resource Language: Google AutoML vs. from-scratch NMT Systems. *Translating and the Computer*, 41, 113-124.

Stahlberg, Felix (2020). Neural Machine Translation: A Review. *Journal of Artificial Intelligence Research*, 69, 343-418.

Stanovsky, Gabriel; Smith, Noah A. & Zettlemoyer, Luke (2019). Evaluating Gender Bias in Machine Translation. *CoRR abs/1906.00591*.

Sutskever, Ilya; Vinyals, Oriol & Le, Quoc V. (2014). Sequence to Sequence Learning with Neural Networks. *Advances in neural information processing systems*, 27, 1-9.

Tan, Zhixing; Wang, Shuo; Yang, Zonghan; Chen, Gang; Huang, X Xuancheng; Sun, Maosong & Liu, Yang (2020). Neural machine translation: A review of methods, resources, and tools. *AI Open*, 1, 5-21.

Tezcan, Arda; Hoste, Véronique & Macken, Lieve (2017). SCATE Taxonomy and Corpus of Machine Translation Errors. In *Trends in E-Tools and Resources for Translators and Interpreters*, 219-244.

Thouin, Benoit (1981). The METEO system. In *Translating and the Computer: Practical Experience of Machine Translation*, 39-44.

Toma, P. (1977). Systran as a Multilingual Machine Translation System. In *Proceedings of the Third European Congress on Information Systems and Networks, Overcoming the Language Barrier*, 569-581.

Tripathi, Sneha & Sarkhel, Juran Krishna (2010). Approaches to machine translation. *Annals of Library and Information Studies*, vol. 57, 388-393.

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz & Polosukhin, Illia (2017). Attention Is All You Need. *Advances in neural information processing systems*, 30.

Vieira, Lucas N. (2020). Automation Anxiety and Translators. *Translation Studies*, 13:1, 1-21.

Way, Andy (2018). Quality Expectations of Machine Translation. In: Moorkens. J.; Castilho, S.; Gaspari, F. & Doherty S. (eds) *Translation Quality Assessment. Machine Translation: Technologies and Applications*, vol. 1. Cham: Springer, 159-178.

Way, Andy (2020). Machine Translation: Where are we at today? In: Angelone, E.; Massey, G. & Ehrensberger-Dow, M. (eds.) *The Bloomsbury Companion to Language Industry Studies*. London: Bloomsbury, 311-332.

Weaver, Warren (1952). Translation. In *Proceedings of the Conference on Mechanical Translation*.

Wiesmann, Eva (2019). Machine Translation in the Field of Law: A Study of the Translation of Italian Legal Texts into German. *Comparative Legilinguistics*, vol. 37, 117-153.

Williams, Philip; Sennrich, Rico; Post, Matt & Koehn, Philipp (2016). Syntax-based Statistical Machine Translation. *Synthesis Lectures on Human Language Technologies*, 9(4), 1-208.

Zanon, Heinz (2001). Spurensuche 1999: Die deutsche Sprache bei Gericht in Südtirol. In: Egger, K. (ed.) *Die deutsche Sprache in Südtirol. Einheitssprache und regionale Vielfalt,* Vienna: Folio, 168–186.

Zaretskaya, Anna; Pastor, Gloria C. & Seghiri, Miriam (2015). Translators' Requirements for Translation Technologies: a User Survey. *New Horizons in Translation and Interpreting Studies*, 133-134.

Zoph, Barrett; Yuret, Deniz; May, Jonathan & Knight, Kevin (2016). Transfer Learning for Low-Resource Neural Machine Translation. *CoRR abs/1604.02201*.

Web Sources

ASTAT (2021). Südtirol in Zahlen – Alto Adige in cifre. https://astat.provincia.bz.it/downloads/Siz_2021(13).pdf (Last accessed: 30.07.2022).

Bistro (n.d.). http://bistrosearch.eurac.edu/ (Last accessed: 23.07.2022).

Eurac Research (2022a). SSL-Laien – Optimizing expert-to-lay communication. Case study: e-learning modules of the Autonomous Province of Bolzano - South Tyrol. https://www.eurac.edu/en/institutes-centers/institute-for-applied-linguistics/projects/ssl-laien (Last accessed: 30.05.2022).

Eurac Research (2022b). SSL-Eurac – Texts on workplace safety and gender-neutral language in Eurac Research. https://www.eurac.edu/en/institutes-centers/institute-for-applied-linguistics/projects/ssl-eurac (Last accessed: 30.05.2022).

INAIL (2018). ANTRAGSFORMULAR für die Reduzierung des mittleren Prämiensatzes im Sinne des Art. 24 der Anwendungsbestimmungen der Prämientarife (M.D. 12/12/2000 und nachfolgende Abänderungen) nach den ersten zwei Jahren Tätigkeit. https://www.inail.it/cs/internet/docs/mod-ot24-tedesco.pdf?section=atti-e-documenti (Last accessed: 29.07.2022).

INAIL (2022). Infortuni e malattie professionali, online gli open data mensili Inail del 2021. https://www.inail.it/cs/internet/comunicazione/sala-stampa/comunicati-stampa/com-stampa-open-data-2021.html (Last accessed: 29.07.2022).

LexBrowser (n.d.). http://lexbrowser.provinz.bz.it/it (Last accessed: 24.05.2022).

ModernMT (n.d.) Translate. https://www.modernmt.com/translate/ (Last accessed: 23.07.2022).

OPUS (n.d). OPUS … the open parallel corpus. https://opus.nlpl.eu (Last accessed: 22.07.2022).

9. Appendix

In the following three sub-sections of the appendix, the source text created by the Institute for Applied Linguistics of Eurac Research together with the translations performed by the adapted and the baseline models are given. This test set consisting of 79 segments also served for the human evaluation of the adapted model. The segments are numbered to facilitate the comparison between translations.

9.1. Source text

1. Il diritto al lavoro è sancito dalla Costituzione Italiana.
2. A partire dagli anni 1950, la normativa italiana ed europea hanno progressivamente accolto i concetti di salute e di sicurezza sul lavoro. La salute è intesa come benessere fisico, mentale e sociale e non solo come assenza di malattia.
3. Ognuno gode quindi del diritto a svolgere un lavoro SICURO, che elimini o comunque minimizzi i rischi per la propria salute e sicurezza.
4. È importante proteggere sempre chi lavora da infortuni o morte (sicurezza) o malattie (salute) causati dal lavoro, secondo le migliori misure tecniche, organizzative e gestionali disponibili in quel momento.
5. Una malattia è considerata professionale quando sussiste un nesso causale tra il lavoro e il rischio che ha causato la malattia. Il rischio può essere provocato dall'attività lavorativa oppure dall'ambiente in cui si svolge il lavoro.
6. Chi lavora è più esposto alle cause della patologia rispetto al resto della popolazione. Tra le possibili cause diluite nel tempo si conosce, ad esempio, l'esposizione a: sostanze chimiche, agenti biologici, rumore, vibrazioni, radiazioni, videoterminali.
7. L'assicurazione INAIL copre i danni provocati dalle malattie professionali prevedendo prestazioni di carattere economico, sanitario e riabilitativo.
8. Le malattie professionali si prevengono con una corretta valutazione dei rischi e con misure di prevenzione e protezione adeguate. Si monitorano tramite la sorveglianza sanitaria effettuata dal medico o dalla medica competente. Se rileva una malattia professionale, il medico o la medica competente deve segnalarla a INAIL.
9. Importante: senza i necessari dispositivi di protezione individuale le attività di laboratorio o il lavoro all'aperto non sono possibili, il loro costo deve essere quindi considerato come parte integrante del relativo budget
10. Le misure di prevenzione e protezione definite dal datore di lavoro in collaborazione con il Servizio di prevenzione e protezione devono essere applicate dall'intera organizzazione.
11. È necessario definire un sistema di supervisione e controllo con chiare responsabilità.
12. Il personale ha un ruolo attivo e segnala a preposti e preposte, dirigenti, datore di lavoro e al Servizio di prevenzione e protezione qualsiasi problema relativo alla salute e sicurezza.
13. Qualsiasi luogo di lavoro presenta rischi per la salute e la sicurezza delle persone, in misura maggiore o minore a seconda dell'attività svolta (un ufficio presenta chiaramente rischi ridotti rispetto a un laboratorio chimico).
14. La valutazione dei rischi corrisponde perciò a un'approfondita analisi di tutte le caratteristiche dei luoghi di lavoro e delle attività che vi si svolgono.

15. La valutazione di tutti i rischi, i loro pesi e le misure di prevenzione e protezione ritenute necessarie sono raccolti all'interno del documento di valutazione dei rischi.
16. L'evento scatenante è improvviso e violento e si differenzia dalle cause della malattia professionale, che sono invece progressive e diluite nel tempo. Tra le possibili cause violente ci sono, ad esempio: parti meccaniche in movimento, sostanze tossiche, sforzi muscolari, condizioni climatiche.
17. L'assicurazione obbligatoria INAIL tutela le persone assicurate contro i danni fisici ed economici che derivano da infortuni causati dall'attività lavorativa o da malattie professionali. Con l'assicurazione il datore o la datrice di lavoro è esonerato dalla responsabilità civile, tranne nei casi in cui sia riconosciuta la sua responsabilità per aver violato le norme sulla salute e sicurezza sul lavoro.
18. Tutti gli infortuni sul lavoro con prognosi di almeno 1 giorno devono poi essere comunicati all'INAIL dal datore di lavoro di Eurac Research.
19. È quindi fondamentale che tutti i quasi-infortuni siano segnalati al Servizio di prevenzione e protezione, così può analizzarli e apportare i necessari cambiamenti affinché simili eventi (ed eventi più gravi) non si ripetano in futuro.
20. Il datore di lavoro di Eurac Research ha l'obbligo di valutare e documentare tutti i rischi per la salute e sicurezza del personale e di identificare adeguate misure di prevenzione e protezione per minimizzarli o eliminarli.
21. Le misure di prevenzione e protezione possono essere tecniche, organizzative o personali.
22. Esempi di misure tecniche: sostituzione di un macchinario non sicuro, manutenzione, aggiunta di dispositivi di protezione collettiva, uso di sostanze meno pericolose.
23. Esempi di misure organizzative: procedure di lavoro sicure, responsabilità e controllo, sorveglianza sanitaria.
24. Esempi di misure personali: informazione, formazione, addestramento, dispositivi di protezione individuale.
25. Una valutazione dei rischi, anche se approfondita, non è sufficiente a garantire salute e sicurezza sul lavoro quando non è seguita da misure di prevenzione e protezione efficaci e da un sistema di gestione interno all'organizzazione che si preoccupi di applicarle e farle osservare in tempi certi.
26. Il personale deve essere consapevole dei rischi per la salute e sicurezza rilevati e conoscere le procedure per eliminarli o minimizzarli. Ciò vale anche per eventuali ospiti.
27. Tutto il personale è considerato videoterminalista e sottoposto a visite mediche periodiche, come da protocollo sanitario specifico elaborato dal medico competente in base ai rischi per la salute rilevati.
28. La selezione di macchinari, attrezzature e strumenti di lavoro che garantiscano salute e sicurezza rientra nella progettazione corretta degli ambienti di lavoro e delle attività lavorative.
29. Il Servizio di prevenzione e protezione si accerta che le manutenzioni periodiche siano effettuate come da normativa.
30. Una manutenzione regolare e adeguata di impianti, macchinari, attrezzature e strumenti di lavoro è di fondamentale importanza per prevenire i rischi per la salute e sicurezza del personale.
31. La selezione di sostanze e preparati chimici, biologici, cancerogeni e mutageni che minimizzino i rischi per salute e sicurezza del personale rientra nella progettazione corretta di ambienti e attività lavorative.
32. La scheda di sicurezza fornisce tutte le istruzioni per usare e smaltire una sostanza in sicurezza. Le indicazioni fornite devono quindi essere sempre rispettate.
33. DATORE O DATRICE DI LAVORO (DL)

34. Definizione: soggetto titolare del rapporto di lavoro con il lavoratore o la lavoratrice
35. Sanzioni massime per inadempienza: arresto fino a 18 mesi o ammenda fino a 15.000 €. A queste si possono aggiungere sanzioni amministrative o pene detentive previste dal codice civile e penale.
36. Il ruolo di datore di lavoro per Eurac Research è attribuito al direttore Stephan Ortner.
37. MEDICO O MEDICA COMPETENTE
38. Definizione: professionista in medicina in possesso di uno dei titoli e dei requisiti formativi e professionali necessari, che collabora con il datore o la datrice di lavoro ai fini della valutazione dei rischi ed è nominato/a dal datore o dalla datrice di lavoro per effettuare la sorveglianza sanitaria e altri compiti relativi alla salute del personale
39. Principali responsabilità:
40. * collabora con il Servizio di prevenzione e protezione alla valutazione dei rischi e alla definizione delle necessarie misure di prevenzione e protezione
41. * programma la sorveglianza sanitaria per il personale in funzione dei rischi rilevati, definendo i protocolli sanitari in base agli indirizzi scientifici più avanzati
42. * verifica l'idoneità alla mansione prevista per ciascun lavoratore o lavoratrice
43. * fornisce informazioni al personale sul significato e sui risultati della sorveglianza sanitaria
44. * partecipa alla redazione dei programmi di informazione e formazione su eventuali rischi per la salute
45. * partecipa all'organizzazione del servizio di primo soccorso
46. Gli istituti di ricerca e center senza infrastrutture di laboratorio e che non svolgono lavoro all'aperto, così come i dipartimenti amministrativi, svolgono attività tendenzialmente ripetitive. In questo caso i rischi sono più facili da valutare e possono essere gestiti con misure precauzionali standard. Secondo una logica che mira a massimizzare la conoscenza della situazione e quindi l'efficacia delle misure di sicurezza, si predilige la massima vicinanza tra la realtà da gestire e chi la gestisce.
47. Lo stesso vale per chi ricopre il ruolo in un istituto di ricerca o center con infrastrutture di laboratorio o che svolge lavoro all'aperto, nei quali la gestione dell'incertezza legata all'attività di ricerca prevede però sempre un dirigente o una dirigente di riferimento.
48. Resta di fondamentale importanza che il Servizio di prevenzione e protezione sia sempre coinvolto nelle decisioni che possono avere un impatto sulla salute e sicurezza del personale (nuove infrastrutture o attrezzature, nuovi progetti o test, nuove sostanze o materiali, nuove attività). Spesso una valutazione dei rischi preventiva garantisce migliori livelli di salute e sicurezza e riduce i costi della messa in sicurezza.
49. 3. LA VALUTAZIONE DEI RISCHI
50. Definizione: valutazione globale e documentata di tutti i rischi per la salute e sicurezza dei lavoratori e delle lavoratrici presenti nell'ambito dell'organizzazione in cui prestano la propria attività, finalizzata ad individuare le adeguate misure di prevenzione e di protezione e ad elaborare il programma delle misure atte a garantire il miglioramento nel tempo dei livelli di salute e sicurezza
51. L'obbligo fondamentale – tanto da non essere mai delegabile – del datore o della datrice di lavoro è la valutazione dei rischi.
52. La valutazione deve riguardare tutti i rischi per la sicurezza e la salute del personale. Riguarda anche la scelta delle attrezzature di lavoro, delle sostanze e dei preparati chimici impiegati, della sistemazione dei luoghi di lavoro. Comprende inoltre i rischi per le lavoratrici in stato di gravidanza e per le persone esposte a rischi particolari, inclusi quelli legati allo stress lavoro-correlato, alle differenze di genere, all'età, alla provenienza da altri Paesi.
53. QUASI-INFORTUNI

54. Un quasi-infortunio (o mancato infortunio, near-miss in inglese) è un qualsiasi evento, legato al lavoro, che avrebbe potuto causare un infortunio o danno alla salute (malattia) o morte ma, solo per puro caso, non lo ha prodotto: un evento quindi che ha in sé la potenzialità di produrre un infortunio. I quasi-infortuni, proporzionalmente molto più numerosi degli infortuni, vanno quindi considerati, come gli infortuni veri e propri, indicatori di rischio. Rappresentano dunque una miniera di informazioni per la prevenzione degli infortuni.
55. Esempi di quasi-infortuni:
56. * collasso di uno scaffale senza conseguenze
57. * caduta di un oggetto dall'alto che non colpisce nessuno
58. * annerimento di arredi o strumentazioni elettriche
59. * scivolamento su di una superficie liscia che non ha però portato a una caduta
60. Questi eventi possono essere precursori di incidenti più gravi. Un'attenta investigazione può fornire indicazioni efficaci sui possibili interventi migliorativi, tecnici e/o gestionali.
61. Lavoratori e lavoratrici devono essere incoraggiati a condividere questi episodi, anche se sono frutto di un errore. Così facendo si ha un'occasione di miglioramento e si potrà evitare che episodi simili si ripetano.
62. Il protocollo sanitario è applicato in base ai profili di rischio attribuiti alla singola persona.
63. L'organizzazione delle visite mediche rientra tra i compiti svolti dal dipartimento Human Resources, che invita il personale alla visita medica. Inoltre, prima della visita comunica al medico competente i profili di rischio del personale da visitare.
64. Ti devi quindi sempre accertare (come i dirigenti e le dirigenti) che il Servizio di prevenzione e protezione sia stato PREVENTIVAMENTE informato dell'introduzione di nuove sostanze. La valutazione dei rischi e la definizione delle misure di prevenzione e protezione necessarie DEVONO INFATTI PRECEDERE L'INTRODUZIONE DI NUOVE SOSTANZE.
65. È tuo compito:
66. * verificare che le sostanze in uso siano le meno pericolose a disposizione per quel determinato utilizzo; quando possibile, sostituire le sostanze con altre meno pericolose; segnalare la possibilità di sostituzione a dirigente, datore di lavoro e al Servizio di prevenzione e protezione
67. * segnalare a dirigente, datore di lavoro e al Servizio di prevenzione e protezione qualsiasi situazione di potenziale rischio legata all'uso di sostanze chimiche, biologiche, cancerogene e mutagene
68. * conoscere le sostanze chimiche, biologiche, cancerogene e mutagene
69. * saper leggere e consultare le schede di sicurezza
70. * identificare un luogo accessibile al personale per conservare la documentazione (es. schede di sicurezza)
71. * informare, formare e addestrare il personale all'uso delle sostanze chimiche, biologiche, cancerogene e mutagene di tua competenza o comunque verificare che sia stato informato, formato e addestrato
72. * verificare che durante l'uso di sostanze chimiche, biologiche, cancerogene e mutagene siano disponibili e impiegati i dispositivi di protezione individuale e le procedure di lavoro e di immagazzinamento o smaltimento indicati dal Servizio di prevenzione e protezione
73. * saper gestire le situazioni di emergenza e istruire il personale alla loro gestione
74. Un infortunio sul lavoro è dunque un infortunio per cui sussiste un nesso causale tra il lavoro e il rischio che lo ha causato. Non è sufficiente che avvenga SUL lavoro (sul luogo di lavoro o durante l'orario di lavoro) ma deve avvenire PER il lavoro. Deve quindi

esistere un rapporto di causa-effetto tra l'attività svolta dalla persona infortunata e l'incidente che ha portato all'infortunio.

75. Tu e/o il personale dovete segnalare qualsiasi infortunio, anche di minima gravità, al dipartimento Human Resources (per la gestione amministrativa dell'infortunio) e al Servizio di prevenzione e protezione (per un approfondimento delle cause che lo hanno generato).
76. SELEZIONE E USO CORRETTI DI IMPIANTI, MACCHINARI, ATTREZZATURE E STRUMENTI DI LAVORO
77. Ti devi quindi sempre accertare (come anche i dirigenti e le dirigenti) che il Servizio di prevenzione e protezione sia stato PREVENTIVAMENTE informato dell'introduzione di nuovi macchinari, attrezzature o strumenti di lavoro. In questo modo il Servizio può valutare i rischi e definire le misure di prevenzione e protezione necessarie PRIMA CHE AVVENGA LA MODIFICA.
78. RESPONSABILE DEL SERVIZIO DI PREVENZIONE E PROTEZIONE (RSPP)
79. Definizione: persona in possesso delle capacità e dei necessari requisiti professionali designata dal datore o dalla datrice di lavoro, a cui risponde, per coordinare il Servizio di prevenzione e protezione dai rischi.

9.2. Translation performed by the adapted model

1. Das Recht auf Arbeit ist in der italienischen Verfassung verankert.
2. Seit den 1950er Jahren haben die italienischen und europäischen Rechtsvorschriften zunehmend die Konzepte der Gesundheit und Sicherheit am Arbeitsplatz übernommen. Gesundheit wird als körperliches, geistiges und soziales Wohlbefinden verstanden und nicht nur als Abwesenheit von Krankheit.
3. Jeder hat somit das Recht auf eine SICHERE Arbeit, die die Risiken für die eigene Gesundheit und Sicherheit beseitigt oder minimiert.
4. Es ist immer wichtig, die Arbeitnehmer vor Arbeitsunfällen oder -tod (Sicherheit) oder arbeitsbedingten Krankheiten (Gesundheit) zu schützen, und zwar gemäß den besten technischen, organisatorischen und betrieblichen Maßnahmen, die zu diesem Zeitpunkt verfügbar sind.
5. Eine Krankheit gilt als Berufskrankheit, wenn ein ursächlicher Zusammenhang zwischen der Arbeit und dem Risiko besteht, das die Krankheit verursacht hat. Die Gefahr kann durch die Arbeitstätigkeit oder das Arbeitsumfeld verursacht werden.
6. Wer arbeitet, ist den Ursachen der Krankheit stärker ausgesetzt als der Rest der Bevölkerung. Mögliche Ursachen, die im Laufe der Zeit verdünnt werden können, sind z. B. die Exposition gegenüber: Chemikalien, biologischen Stoffen, Lärm, Vibrationen, Strahlung und Videoterminals.
7. Die INAIL-Versicherung deckt die durch Berufskrankheiten verursachten Schäden durch wirtschaftliche, gesundheitliche und rehabilitative Leistungen ab.
8. Berufskrankheiten werden durch eine korrekte Risikobewertung und entsprechende Arbeitsschutzmaßnahmen verhindert. Sie werden durch die ärztliche Überwachung durch den zuständigen Arzt oder die zuständige Ärztin überwacht. Wenn Sie eine Berufskrankheit feststellen, muss Ihr Arzt oder Ihre Ärztin dies INAIL melden.
9. Wichtig: Ohne die nötige persönliche Schutzausrüstung sind Laboraufgaben oder Arbeiten im Freien nicht möglich, deren Kosten daher als Bestandteil des Budgets zu betrachten sind.
10. Die vom Arbeitgeber in Zusammenarbeit mit dem Arbeitsschutzdienst festgelegten Arbeitsschutzmaßnahmen müssen von der gesamten Organisation umgesetzt werden.

11. Es muss ein Überwachungs- und Kontrollsystem mit klarer Verantwortung festgelegt werden.
12. Das Personal nimmt eine aktive Rolle ein und meldet den Vorgesetzten, Führungskräften, Arbeitgebern und dem Arbeitsschutzdienst alle Fragen im Zusammenhang mit Gesundheit und Sicherheit.
13. Jeder Arbeitsplatz birgt je nach Tätigkeit mehr oder weniger Risiken für die Gesundheit und Sicherheit von Menschen (ein Büro birgt eindeutig geringere Risiken als ein chemisches Labor).
14. Die Risikobewertung entspricht daher einer eingehenden Analyse aller Merkmale der Arbeitsplätze und der dort ausgeübten Tätigkeiten.
15. Die Bewertung aller Risiken, deren Gewicht und die als notwendig erachteten Arbeitsschutzmaßnahmen werden im Risikobewertungsdokument zusammengefasst.
16. Das auslösende Ereignis ist plötzlich und gewalttätig und unterscheidet sich von den Ursachen der Berufskrankheit, die im Laufe der Zeit progressiv und verdünnt sind. Mögliche gewaltsame Ursachen sind z. B. bewegte Maschinenteile, toxische Stoffe, Muskelanstrengungen, Witterungsverhältnisse.
17. Die obligatorische INAIL-Versicherung schützt versicherte Personen vor körperlichen und wirtschaftlichen Schäden, die durch arbeitsbedingte Unfälle oder Berufskrankheiten entstehen. Mit der Versicherung ist der Arbeitgeber oder die Arbeitgeberin von der zivilrechtlichen Haftung befreit, es sei denn, seine Haftung für die Verletzung der Gesundheits- und Sicherheitsvorschriften am Arbeitsplatz wird anerkannt.
18. Alle Arbeitsunfälle mit einer Prognose von mindestens 1 Tag müssen dem INAIL vom Arbeitgeber von Eurac Research gemeldet werden.
19. Es ist daher wichtig, dass alle Beinaheunfälle dem Arbeitsschutzdienst gemeldet werden, damit dieser sie analysieren und die notwendigen Änderungen vornehmen kann, damit sich solche Ereignisse (und schwerwiegendere Ereignisse) in Zukunft nicht wiederholen.
20. Der Arbeitgeber von Eurac Research ist verpflichtet, alle Risiken für die Gesundheit und Sicherheit des Personals zu bewerten und zu dokumentieren und geeignete Präventions- und Schutzmaßnahmen zu identifizieren, um sie zu minimieren oder zu beseitigen.
21. Arbeitsschutzmaßnahmen können technische, organisatorische oder persönliche Maßnahmen sein.
22. Beispiele für technische Maßnahmen: Austausch unsicherer Maschinen, Wartung, Zusatz von kollektiven Schutzausrüstungen, Verwendung weniger gefährlicher Stoffe.
23. Beispiele für organisatorische Maßnahmen: sichere Arbeitsabläufe, Verantwortung und Kontrolle, Gesundheitsüberwachung.
24. Beispiele für persönliche Maßnahmen: Information, Schulung, Schulung, persönliche Schutzausrüstungen.
25. Eine gründliche Risikobewertung reicht nicht aus, um die Sicherheit und den Gesundheitsschutz am Arbeitsplatz zu gewährleisten, wenn sie nicht von wirksamen Arbeitsschutzmaßnahmen und einem organisationsinternen Managementsystem begleitet wird, das sich darum kümmert, sie rechtzeitig umzusetzen und durchzusetzen.
26. Das Personal muss sich der festgestellten Gesundheits- und Sicherheitsrisiken bewusst sein und die Verfahren zur Beseitigung oder Minimierung dieser Risiken kennen. Das gilt auch für eventuelle Gäste.
27. Alle Mitarbeiter gelten als Videoterminalisten und werden regelmäßig medizinisch untersucht, gemäß dem spezifischen Gesundheitsprotokoll, das vom zuständigen Arzt auf der Grundlage der festgestellten Gesundheitsrisiken erstellt wird.
28. Die Auswahl von Maschinen, Geräten und Arbeitsmitteln, die Gesundheit und Sicherheit gewährleisten, ist Teil der richtigen Gestaltung der Arbeitsplätze und der Arbeitstätigkeiten.

29. Der Dienst für Arbeitsschutz stellt sicher, dass die regelmäßigen Wartungsarbeiten gemäß den Vorschriften durchgeführt werden.
30. Die regelmäßige und ordnungsgemäße Wartung von Anlagen, Maschinen, Geräten und Arbeitsmitteln ist von grundlegender Bedeutung, um Risiken für die Gesundheit und Sicherheit des Personals zu vermeiden.
31. Die Auswahl von chemischen, biologischen, krebserregenden und erbgutverändernden Stoffen und Zubereitungen, die die Risiken für die Gesundheit und Sicherheit des Personals minimieren, ist Teil der korrekten Gestaltung von Arbeitsumgebungen und -aktivitäten.
32. Das Sicherheitsdatenblatt enthält alle Anweisungen für die sichere Verwendung und Entsorgung eines Stoffes. Die vorgegebenen Hinweise sind daher stets einzuhalten.
33. ARBEITGEBER ODER ARBEITGEBERIN (DL)
34. Definition: Träger des Arbeitsverhältnisses mit dem Arbeitnehmer oder der Arbeitnehmerin
35. Höchststrafe bei Nichteinhaltung: Haftstrafe bis zu 18 Monaten oder Geldbuße bis zu 15.000 €. Dazu können verwaltungsrechtliche Sanktionen oder im Zivil- und Strafgesetzbuch vorgesehene Gefängnisstrafen hinzukommen.
36. Die Rolle des Arbeitgebers für Eurac Research wird dem Direktor Stephan Ortner zugewiesen.
37. BETRIEBSARZT ODER BETRIEBSARZT
38. Definition: Mediziner, der über einen der erforderlichen Titel, beruflichen Befähigungen und Ausbildungen verfügt, mit dem Arbeitgeber oder der Arbeitgeberin zusammenarbeitet, um die Risikobewertung durchzuführen, und der vom Arbeitgeber oder der Arbeitgeberin mit der Durchführung der Gesundheitsüberwachung und anderen mit der Gesundheit des Personals zusammenhängenden Aufgaben betraut ist.
39. Hauptaufgaben:
40. * arbeitet mit dem Arbeitsschutzdienst bei der Risikobewertung und bei der Festlegung der notwendigen Arbeitsschutzmaßnahmen zusammen
41. * plant die Gesundheitsüberwachung für das Personal entsprechend den festgestellten Risiken und definiert die Gesundheitsprotokolle auf der Grundlage der fortschrittlichsten wissenschaftlichen Richtlinien
42. * Überprüfung der Eignung für die für jeden Arbeitnehmer oder jede Arbeitnehmerin vorgesehene Aufgabe
43. * informiert das Personal über die Bedeutung und die Ergebnisse der Gesundheitsüberwachung
44. * beteiligt sich an der Ausarbeitung von Informations- und Ausbildungsprogrammen über mögliche Gesundheitsrisiken
45. * beteiligt sich an der Organisation des Erste-Hilfe-Dienstes
46. Forschungsinstitute und Zentren ohne Laborinfrastrukturen, die nicht im Freien arbeiten, sowie Verwaltungsabteilungen führen tendenziell sich wiederholende Aktivitäten durch. In diesem Fall sind die Risiken leichter einzuschätzen und können mit Standardvorkehrungen gemanagt werden. Nach einer Logik, die darauf abzielt, die Kenntnis der Situation und damit die Wirksamkeit der Sicherheitsmaßnahmen zu maximieren, wird die maximale Nähe zwischen der zu verwaltenden Realität und demjenigen, der sie verwaltet, bevorzugt.
47. Dasselbe gilt für diejenigen, die in einer Forschungseinrichtung oder einem Zentrum mit Laborinfrastruktur tätig sind oder im Freien arbeiten, in denen das Management der Unsicherheit, die mit der Forschungstätigkeit verbunden ist, jedoch immer einen Manager oder eine Referenzleiterin erfordert.

48. Es ist von entscheidender Bedeutung, dass der Arbeitsschutzdienst stets an Entscheidungen beteiligt wird, die sich auf die Gesundheit und Sicherheit des Personals auswirken können (neue Infrastrukturen oder Ausrüstungen, neue Projekte oder Tests, neue Stoffe oder Materialien, neue Tätigkeiten). Eine präventive Risikobewertung sorgt häufig für ein höheres Gesundheits- und Sicherheitsniveau und senkt die Kosten für die Sicherheit.
49. 3. DIE RISIKOBEWERTUNG
50. Definition: dokumentierte Gesamtbewertung aller Risiken für die Gesundheit und Sicherheit der Arbeitnehmer in der Organisation, in der sie ihre Arbeit leisten, welche darauf ausgerichtet ist, die angemessenen Präventions- und Schutzmaßnahmen zu finden und ein Maßnahmenprogramm auszuarbeiten, um eine Verbesserung des Gesundheits- und Sicherheitsniveaus langfristig zu garantieren
51. Die grundlegende Verpflichtung – die niemals delegierbar ist – des Arbeitgebers oder der Arbeitgeberin ist die Risikobewertung.
52. Die Bewertung muss alle Risiken für die Sicherheit und die Gesundheit des Personals abdecken. Es geht auch um die Auswahl der Arbeitsmittel, der verwendeten chemischen Stoffe und Zubereitungen und der Arbeitsstätten. Es umfasst auch Risiken für schwangere Arbeitnehmerinnen und für Personen, die besonderen Risiken ausgesetzt sind, einschließlich arbeitsbedingtem Stress, geschlechtsspezifischen Unterschieden, Alter, Herkunft aus anderen Ländern.
53. QUASI-INFORTUNI
54. Ein Beinahe-Unfall (oder Near-Miss auf Englisch) ist ein arbeitsbedingtes Ereignis, das einen Unfall oder Gesundheitsschaden (Krankheit) oder Tod verursacht haben könnte, aber nur durch reinen Zufall nicht verursacht hat: ein Ereignis, das an sich das Potenzial hat, einen Unfall zu verursachen. Die Quasi-Unfälle, die proportional viel zahlreicher sind als die Unfälle, sind daher wie die eigentlichen Unfälle als Risikoindikatoren zu betrachten. Sie stellen somit eine Informationsquelle für die Unfallprävention dar.
55. Beispiele von Beinahe-Unfällen:
56. * Zusammenbruch eines Regals ohne Folgen
57. * Absturz eines Gegenstandes von oben, der niemanden trifft
58. * Schwärzung von Einrichtungsgegenständen oder elektrischen Geräten
59. * Rutschen auf einer glatten Oberfläche, die jedoch nicht zu einem Sturz geführt hat
60. Diese Ereignisse können Vorboten schwerer Unfälle sein. Eine sorgfältige Untersuchung kann wirksame Hinweise auf mögliche Verbesserungs-, technische und/oder Managementmaßnahmen geben.
61. Arbeitnehmerinnen und Arbeitnehmer müssen ermutigt werden, diese Vorfälle zu teilen, auch wenn sie das Ergebnis eines Fehlers sind. Dadurch haben Sie eine Chance, sich zu verbessern, und Sie können verhindern, dass sich ähnliche Episoden wiederholen.
62. Das Gesundheitsprotokoll wird auf der Grundlage der der einzelnen Person zugeordneten Risikoprofile angewendet.
63. Die Organisation von Arztbesuchen gehört zu den Aufgaben der Abteilung Human Resources, die das Personal zur ärztlichen Untersuchung einlädt. Darüber hinaus teilt er dem zuständigen Arzt vor dem Besuch die Risikoprofile des zu besuchenden Personals mit.
64. Stellen Sie daher immer sicher (wie Führungskräfte), dass der Arbeitsschutzdienst IM VORAUS über die Einführung neuer Stoffe informiert wurde. Die Risikobewertung und die Festlegung der erforderlichen Arbeitsschutzmaßnahmen MÜSSEN VOR DER Einführung NEUER STOFFE erfolgen.
65. Es ist Ihre Aufgabe:

66. * Überprüfen Sie, ob die verwendeten Substanzen für diese Verwendung am wenigsten gefährlich sind; wenn möglich, ersetzen Sie die Substanzen durch andere, die weniger gefährlich sind; melden Sie die Möglichkeit des Austauschs an Führungskräfte, Arbeitgeber und den Arbeitsschutzdienst
67. * Unterrichtung des Managements, des Arbeitgebers und des Arbeitsschutzdienstes über alle potenziellen Risikosituationen im Zusammenhang mit der Verwendung von chemischen, biologischen, krebserzeugenden und erbgutverändernden Stoffen
68. * die chemischen, biologischen, krebserregenden und erbgutverändernden Stoffe kennen
69. * Sicherheitsdatenblätter lesen und konsultieren können
70. * Identifizieren Sie einen Ort, der für das Personal zugänglich ist, um die Unterlagen aufzubewahren (z. B. Sicherheitsdatenblätter)
71. * das Personal über die Verwendung der chemischen, biologischen, krebserregenden und erbgutverändernden Stoffe, für die Sie zuständig sind, informieren, schulen und schulen oder anderweitig sicherstellen, dass es informiert, geschult und geschult wurde
72. * Sicherstellen, dass bei der Verwendung von chemischen, biologischen, krebserregenden und erbgutverändernden Stoffen die von der Dienststelle für Prävention und Schutz angegebenen persönlichen Schutzausrüstungen sowie die Arbeits- und Lagerungs- oder Entsorgungsverfahren zur Verfügung stehen und verwendet werden
73. * in der Lage sein, mit Notfallsituationen umzugehen und das Personal in deren Bewältigung zu unterweisen
74. Ein Arbeitsunfall ist also ein Unfall, bei dem ein ursächlicher Zusammenhang zwischen der Arbeit und dem Risiko besteht, das sie verursacht hat. Es reicht nicht aus, dass dies bei der Arbeit geschieht (am Arbeitsplatz oder während der Arbeitszeit), sondern es muss BEI der Arbeit geschehen. Es muss also ein kausaler Zusammenhang zwischen der Tätigkeit des Verletzten und dem Unfall, der zum Unfall geführt hat, bestehen.
75. Sie und/oder das Personal müssen alle Verletzungen, auch solche von geringer Schwere, der Abteilung Human Resources (für die Verwaltung des Unfalls) und dem Arbeitsschutzdienst (für eine Vertiefung der Ursachen, die ihn verursacht haben) melden.
76. KORREKTE AUSWAHL UND VERWENDUNG VON ANLAGEN, MASCHINEN, ARBEITSMITTELN UND WERKZEUGEN
77. Sie müssen daher immer sicherstellen, dass der Arbeitsschutzdienst VORAB über die Einführung neuer Arbeitsmaschinen, Arbeitsmittel oder Arbeitsmittel informiert wurde. Auf diese Weise kann der Dienst die Risiken bewerten und die erforderlichen Arbeitsschutzmaßnahmen definieren, BEVOR DIE ÄNDERUNG ERFOLGT.
78. LEITER DES ARBEITSSCHUTZDIENSTES
79. Definition: Person, die über die beruflichen Fähigkeiten und Qualifikationen verfügt, um den Arbeitsschutzdienst zu koordinieren, und die dem Arbeitgeber untersteht.

9.3. Translation performed by the baseline model

1. Das Recht auf Arbeit ist in der italienischen Verfassung verankert.
2. Seit den 1950er Jahren haben die italienischen und europäischen Rechtsvorschriften die Konzepte der Gesundheit und Sicherheit am Arbeitsplatz schrittweise übernommen. Gesundheit wird als körperliches, geistiges und soziales Wohlbefinden verstanden und nicht nur als Abwesenheit von Krankheit.
3. Jeder hat somit das Recht, eine SICHERE Arbeit zu verrichten, die die Risiken für seine Gesundheit und Sicherheit beseitigt oder auf andere Weise minimiert.

4. Es ist immer wichtig, die Arbeitnehmer vor Verletzungen oder Tod (Sicherheit) oder Krankheiten (Gesundheit) zu schützen, die durch die Arbeit verursacht werden, gemäß den besten technischen, organisatorischen und Managementmaßnahmen, die zu diesem Zeitpunkt verfügbar sind.
5. Eine Krankheit gilt als Berufskrankheit, wenn ein ursächlicher Zusammenhang zwischen der Arbeit und dem Risiko besteht, das die Krankheit verursacht hat. Das Risiko kann durch die Arbeitstätigkeit oder die Umgebung, in der die Arbeit ausgeführt wird, verursacht werden.
6. Wer arbeitet, ist den Ursachen der Pathologie stärker ausgesetzt als der Rest der Bevölkerung. Unter den möglichen Ursachen, die im Laufe der Zeit verdünnt werden, ist beispielsweise die Exposition gegenüber: Chemikalien, biologischen Arbeitsstoffen, Lärm, Vibrationen, Strahlung, Videoterminals bekannt.
7. Die INAIL-Versicherung deckt die durch Berufskrankheiten verursachten Schäden durch wirtschaftliche, gesundheitliche und rehabilitative Leistungen ab.
8. Berufskrankheiten werden durch eine korrekte Risikobewertung und geeignete Präventions- und Schutzmassnahmen verhindert. Sie werden durch ärztliche Überwachung durch den zuständigen Arzt oder die zuständige Ärztin überwacht. Wenn Sie eine Berufskrankheit feststellen, muss Ihr Arzt oder die zuständige Ärztin dies INAIL melden.
9. Wichtig: Ohne die notwendige persönliche Schutzausrüstung sind Laboraktivitäten oder Arbeiten im Freien nicht möglich, ihre Kosten müssen daher als integraler Bestandteil des jeweiligen Budgets betrachtet werden.
10. Die vom Arbeitgeber in Zusammenarbeit mit dem Präventions- und Schutzdienst definierten Präventions- und Schutzmassnahmen müssen von der gesamten Organisation umgesetzt werden.
11. Es ist notwendig, ein Überwachungs- und Kontrollsystem mit klarer Verantwortung zu definieren.
12. Das Personal nimmt eine aktive Rolle ein und meldet den Vorgesetzten, Führungskräften, Arbeitgebern und dem Präventions- und Schutzdienst alle Fragen im Zusammenhang mit Gesundheit und Sicherheit.
13. Jeder Arbeitsplatz birgt je nach Tätigkeit mehr oder weniger Risiken für die Gesundheit und Sicherheit von Personen (ein Büro birgt eindeutig geringere Risiken als ein chemisches Labor).
14. Die Risikobewertung entspricht daher einer gründlichen Analyse aller Merkmale der Arbeitsstätten und der dort ausgeübten Tätigkeiten.
15. Die Bewertung aller Risiken, ihre Gewichte und die als notwendig erachteten Präventions- und Schutzmaßnahmen werden im Risikobewertungsdokument zusammengefasst.
16. Das auslösende Ereignis ist plötzlich und gewalttätig und unterscheidet sich von den Ursachen der Berufskrankheit, die im Laufe der Zeit progressiv und verdünnt sind. Mögliche gewalttätige Ursachen sind zum Beispiel: bewegliche mechanische Teile, toxische Substanzen, Muskelanstrengungen, klimatische Bedingungen.
17. Die obligatorische INAIL-Versicherung schützt versicherte Personen vor körperlichen und wirtschaftlichen Schäden, die durch Arbeitsunfälle oder Berufskrankheiten entstehen. Mit der Versicherung ist der Arbeitgeber oder die Arbeitgeberin von der zivilrechtlichen Haftung befreit, es sei denn, seine Haftung für Verstöße gegen die Vorschriften über Sicherheit und Gesundheitsschutz am Arbeitsplatz wird anerkannt.
18. Sämtliche Arbeitsunfälle mit einer Prognose von mindestens einem Tag müssen dem INAIL vom Arbeitgeber von Eurac Research gemeldet werden.

19. Es ist daher wichtig, dass alle Quasi-Unfälle dem Präventions- und Schutzdienst gemeldet werden, damit er sie analysieren und die notwendigen Änderungen vornehmen kann, damit sich solche Ereignisse (und schwerwiegendere Ereignisse) in Zukunft nicht wiederholen.
20. Der Arbeitgeber von Eurac Research ist verpflichtet, alle Risiken für die Gesundheit und Sicherheit des Personals zu bewerten und zu dokumentieren und geeignete Präventions- und Schutzmaßnahmen zu identifizieren, um diese zu minimieren oder zu eliminieren.
21. Präventions- und Schutzmaßnahmen können technischer, organisatorischer oder persönlicher Art sein.
22. Beispiele für technische Maßnahmen: Austausch unsicherer Maschinen, Wartung, Zusatz kollektiver Schutzausrüstung, Verwendung weniger gefährlicher Stoffe.
23. Beispiele für organisatorische Maßnahmen: sichere Arbeitsabläufe, Verantwortung und Kontrolle, Gesundheitsüberwachung.
24. Beispiele für persönliche Maßnahmen: Information, Schulung, Schulung, persönliche Schutzausrüstung.
25. Eine gründliche Risikobewertung reicht nicht aus, um Gesundheit und Sicherheit am Arbeitsplatz zu gewährleisten, wenn sie nicht von wirksamen Präventions- und Schutzmaßnahmen und einem organisationsinternen Managementsystem begleitet wird, das sich darum kümmert, diese rechtzeitig umzusetzen und durchzusetzen.
26. Das Personal muss sich der festgestellten Gesundheits- und Sicherheitsrisiken bewusst sein und die Verfahren zur Beseitigung oder Minimierung dieser Risiken kennen. Dies gilt auch für eventuelle Gäste.
27. Das gesamte Personal gilt als Videoterminalist und wird regelmäßig medizinisch untersucht, gemäß einem spezifischen Gesundheitsprotokoll, das vom zuständigen Arzt auf der Grundlage der festgestellten Gesundheitsrisiken erstellt wird.
28. Die Auswahl von Maschinen, Geräten und Arbeitsmitteln, die Gesundheit und Sicherheit gewährleisten, ist Teil der richtigen Gestaltung von Arbeitsumgebungen und Arbeitsaktivitäten.
29. Der Präventions- und Schutzdienst stellt sicher, dass die regelmäßigen Wartungen gemäß den Vorschriften durchgeführt werden.
30. Regelmäßige und ordnungsgemäße Wartung von Anlagen, Maschinen, Geräten und Arbeitsmitteln ist von grundlegender Bedeutung, um Risiken für die Gesundheit und Sicherheit des Personals zu vermeiden.
31. Die Auswahl chemischer, biologischer, krebserregender und erbgutverändernder Stoffe und Zubereitungen, die die Risiken für die Gesundheit und Sicherheit des Personals minimieren, ist Teil der richtigen Gestaltung von Arbeitsumgebungen und -aktivitäten.
32. Das Sicherheitsdatenblatt enthält alle Anweisungen zur sicheren Verwendung und Entsorgung eines Stoffes. Die vorgelegten Hinweise müssen daher stets eingehalten werden.
33. ARBEITGEBER ODER ARBEITGEBERIN (DL)
34. Definition: Person, die das Arbeitsverhältnis mit dem Arbeitnehmer oder der Arbeitnehmerin
35. Höchststrafen bei Nichteinhaltung: Festnahme bis zu 18 Monaten oder Geldstrafe bis zu 15.000 €. Dazu können verwaltungsrechtliche Sanktionen oder im Zivil- und Strafgesetzbuch vorgesehene Gefängnisstrafen hinzukommen.
36. Die Rolle des Arbeitgebers für Eurac Research wird dem Direktor Stephan Ortner zugewiesen.
37. Kompetente ÄRZTIN oder KOMPETENTE ÄRZTIN
38. Definition: Ärztin/Arzt, die/der über einen der erforderlichen Ausbildungsnachweise und -anforderungen verfügt, bei der Risikobewertung mit dem Arbeitgeber oder der

Arbeitgeberin zusammenarbeitet und vom Arbeitgeber oder der Arbeitgeberin für die Durchführung der Gesundheitsüberwachung und anderer Aufgaben im Zusammenhang mit der Gesundheit des Personals bestellt wird

39. Hauptverantwortung:
40. * arbeitet mit dem Präventions- und Schutzdienst zusammen, um die Risiken zu bewerten und die notwendigen Präventions- und Schutzmaßnahmen festzulegen
41. * Planung der Gesundheitsüberwachung des Personals entsprechend den festgestellten Risiken durch Festlegung von Gesundheitsprotokollen nach den modernsten wissenschaftlichen Leitlinien
42. * Überprüfung der Eignung für die vorgesehene Aufgabe für jeden Arbeitnehmer oder jede Arbeitnehmerin
43. * informiert das Personal über die Bedeutung und die Ergebnisse der Gesundheitsüberwachung
44. * beteiligt sich an der Ausarbeitung von Informations- und Schulungsprogrammen über mögliche Gesundheitsrisiken
45. * beteiligt sich an der Organisation des Erste-Hilfe-Dienstes
46. Forschungsinstitute und Zentren ohne Laborinfrastruktur, die nicht im Freien arbeiten, sowie Verwaltungsabteilungen führen tendenziell sich wiederholende Tätigkeiten durch. In diesem Fall sind die Risiken leichter einzuschätzen und können mit Standardvorkehrungen gehandhabt werden. Nach einer Logik, die darauf abzielt, die Kenntnis der Situation und damit die Wirksamkeit der Sicherheitsmaßnahmen zu maximieren, wird die maximale Nähe zwischen der zu verwaltenden Realität und demjenigen, der sie verwaltet, bevorzugt.
47. Dasselbe gilt für diejenigen, die in einer Forschungseinrichtung oder einem Zentrum mit Laborinfrastruktur tätig sind oder im Freien arbeiten, in denen das Management der Unsicherheit im Zusammenhang mit der Forschungstätigkeit jedoch immer einen Manager oder eine Referenzleiterin vorsieht.
48. Es ist von entscheidender Bedeutung, dass der Präventions- und Schutzdienst stets an Entscheidungen beteiligt wird, die sich auf die Gesundheit und Sicherheit des Personals auswirken können (neue Infrastruktur oder Ausrüstung, neue Projekte oder Tests, neue Stoffe oder Materialien, neue Aktivitäten). Eine vorbeugende Risikobewertung sorgt häufig für ein höheres Gesundheits- und Sicherheitsniveau und senkt die Kosten für die Sicherheit.
49. 3. RISIKOBEWERTUNG
50. Definition: umfassende und dokumentierte Bewertung aller Gesundheits- und Sicherheitsrisiken für Arbeitnehmerinnen und Arbeitnehmer in der Organisation, in der sie tätig sind, mit dem Ziel, geeignete Präventions- und Schutzmaßnahmen zu ermitteln und ein Maßnahmenprogramm zu entwickeln, um die Verbesserung des Gesundheits- und Sicherheitsniveaus im Laufe der Zeit zu gewährleisten
51. Die grundlegende Pflicht – die niemals delegierbar ist – des Arbeitgebers oder der Arbeitgeberin ist die Risikobewertung.
52. Die Bewertung muss sich auf alle Risiken für die Sicherheit und Gesundheit des Personals erstrecken. Es geht auch um die Wahl der Arbeitsmittel, der verwendeten chemischen Stoffe und Zubereitungen sowie um die Anordnung der Arbeitsplätze. Es umfasst auch Risiken für schwangere Arbeitnehmerinnen und für Personen, die besonderen Risiken ausgesetzt sind, einschließlich arbeitsbedingtem Stress, geschlechtsspezifischen Unterschieden, Alter und Herkunft aus anderen Ländern.
53. QUASI-INFORTUNI
54. Eine Quasi-Verletzung (oder eine Nicht-Verletzung, near-miss auf Englisch) ist ein arbeitsbedingtes Ereignis, das eine Verletzung oder einen Gesundheitsschaden

(Krankheit) oder Tod verursacht haben könnte, aber nur durch reinen Zufall nicht verursacht hat: ein Ereignis, das also das Potenzial hat, eine Verletzung hervorzurufen. Quasi-Unfälle, die proportional viel zahlreicher sind als Unfälle, sind daher wie die eigentlichen Unfälle als Risikoindikatoren zu betrachten. Sie stellen somit eine Informationsquelle für die Unfallverhütung dar.

55. Beispiele für Quasi-Verletzungen:
56. * Zusammenbruch eines Regals ohne Folgen
57. * Absturz eines Objekts von oben, das niemanden trifft
58. * Schwärzung von Einrichtungsgegenständen oder elektrischen Geräten
59. * Rutschen auf einer glatten Oberfläche, die jedoch nicht zu einem Sturz geführt hat
60. Diese Ereignisse können Vorboten für schwerwiegendere Unfälle sein. Eine sorgfältige Untersuchung kann wirksame Hinweise auf mögliche Verbesserungs-, technische und/oder Managementmaßnahmen geben.
61. Arbeiterinnen und Arbeiter müssen ermutigt werden, diese Episoden zu teilen, auch wenn sie das Ergebnis eines Fehlers sind. Auf diese Weise haben Sie die Möglichkeit, sich zu verbessern, und Sie können verhindern, dass sich ähnliche Episoden wiederholen.
62. Das Gesundheitsprotokoll wird auf der Grundlage der der einzelnen Person zugeordneten Risikoprofile angewendet.
63. Die Organisation von Arztbesuchen gehört zu den Aufgaben der Abteilung Human Resources, die das Personal zur ärztlichen Untersuchung einlädt. Darüber hinaus teilt er dem zuständigen Arzt vor dem Besuch die Risikoprofile des zu besuchenden Personals mit.
64. Sie müssen daher immer sicherstellen (wie Führungskräfte), dass der Präventions- und Schutzdienst IM VORAUS über die Einführung neuer Substanzen informiert wurde. Die Risikobewertung und die Festlegung der erforderlichen Präventions- und Schutzmaßnahmen MÜSSEN DER Einführung NEUER STOFFE VORAUSGEHEN.
65. Es ist Ihre Aufgabe:
66. * zu überprüfen, ob die verwendeten Substanzen für diese Verwendung am wenigsten gefährlich sind; wenn möglich, die Substanzen durch weniger gefährliche zu ersetzen; die Möglichkeit eines Ersatzes an Führungskräfte, Arbeitgeber und den Präventions- und Schutzdienst zu melden
67. * Unterrichtung des Managements, des Arbeitgebers und des Präventions- und Schutzdienstes über alle potenziellen Risikosituationen im Zusammenhang mit der Verwendung chemischer, biologischer, krebserregender und erbgutverändernder Stoffe
68. * die chemischen, biologischen, krebserregenden und erbgutverändernden Substanzen kennen
69. * Sicherheitsdatenblätter lesen und konsultieren können
70. * Identifizieren Sie einen Ort, der für das Personal zugänglich ist, um die Unterlagen aufzubewahren (z. B. Sicherheitsdatenblätter)
71. * Informieren, schulen und schulen Sie das Personal in Bezug auf die Verwendung der Chemikalien, biologischen, krebserregenden und erbgutverändernden Stoffe, für die Sie zuständig sind, oder vergewissern Sie sich in jedem Fall, dass es informiert, geschult und geschult wurde
72. * Sicherstellen, dass bei der Verwendung von chemischen, biologischen, krebserregenden und erbgutverändernden Stoffen persönliche Schutzausrüstung sowie die vom Präventions- und Schutzdienst angegebenen Arbeits-, Lager- oder Entsorgungsverfahren verfügbar sind und verwendet werden
73. * in der Lage sein, mit Notsituationen umzugehen und das Personal in deren Bewältigung zu unterweisen

74. Ein Arbeitsunfall ist also ein Unfall, bei dem ein ursächlicher Zusammenhang zwischen der Arbeit und dem Risiko besteht, das sie verursacht hat. Es genügt nicht, dass dies bei der Arbeit geschieht (am Arbeitsplatz oder während der Arbeitszeit), sondern es muss BEI der Arbeit geschehen. Es muss also ein kausaler Zusammenhang zwischen der Tätigkeit des Verletzten und dem Unfall, der zum Unfall geführt hat, bestehen.
75. Sie und/oder das Personal müssen alle Verletzungen, auch solche von geringer Schwere, der Abteilung Human Resources (für die Verwaltung des Unfalls) und dem Präventions- und Schutzdienst (zur Vertiefung der Ursachen, die ihn verursacht haben) melden.
76. RICHTIGE AUSWAHL UND VERWENDUNG VON ANLAGEN, MASCHINEN, ARBEITSMITTELN UND WERKZEUGEN
77. Sie müssen daher immer sicherstellen (wie auch Führungskräfte), dass der Präventions- und Schutzdienst IM VORAUS über die Einführung neuer Maschinen, Geräte oder Arbeitswerkzeuge informiert wurde. Auf diese Weise kann der Dienst die Risiken bewerten und die erforderlichen Präventions- und Schutzmaßnahmen festlegen, BEVOR DIE ÄNDERUNG ERFOLGT.
78. LEITER DES PRÄVENTIONS- UND SCHUTZDIENSTES (RSPP)
79. Definition: Person, die über die erforderlichen beruflichen Fähigkeiten und Anforderungen verfügt und vom Arbeitgeber oder der Arbeitgeberin benannt wurde, um den Risikopräventions- und -schutzdienst zu koordinieren.

Abstract

The research field of Machine Translation (MT) has gained increased popularity with the advent of Neural Machine Translation (NMT). Even though modern architectures have reached unexpected quality levels, researchers are trying to solve some common issues to deliver better performances. One of these problems is the inability for MT systems to perform well in a specific domain. The objective of the present research is to test the effects of domain adaptation on an existing NMT system (ModernMT) in the language pair Italian-South Tyrolean German. South Tyrolean German is bound to the Italian legal system and during the years, a specific legal-administrative terminology has been developed, making this language variety particularly challenging to be translated by MT systems. The chosen domain is occupational health and safety, a subdomain of law with very peculiar terminology traits. Occupational health and safety is a central topic in South Tyrol and training courses for employees are written and translated on a yearly basis, making the present master's thesis a valuable attempt to facilitate and accelerate the translation workflow. To fine-tune a pre-trained model, two parallel corpora consisting of roughly 25,000 translation units have been made available by South Tyrolean research center *Eurac Research*. The data have been cleaned, filtered and inserted into ModernMT. The translation quality of three different test sets has been assessed automatically and 79 domain-specific parallel segments have been manually analyzed to annotate accuracy and fluency errors. Results have shown a surprising increase of 6.4 BLEU points and 3.5 METEOR points, as well as a noticeable improvement of the terminological accuracy. However, the presence of multiple terminology errors and mistranslations is the proof that there is still room for improvement.

Abstract in German

Das Forschungsgebiet der maschinellen Übersetzung (MÜ) hat mit dem Aufkommen der neuronalen maschinellen Übersetzung (NMÜ) stark an Popularität gewonnen. Obwohl moderne Architekturen bereits ein erstaunliches Qualitätsniveau erreicht haben, versuchen Forscher weiterhin, einige allgemeine Probleme zu lösen, um bessere Leistungen zu erzielen. Zu den Problemen zählt die Inkompetenz von MÜ-Systemen, in einer bestimmten Domäne gute Leistungen zu erbringen. Ziel der vorliegenden Untersuchung ist es, die Auswirkungen der Domänenanpassung auf ein bestehendes NMÜ-System (ModernMT) im Sprachenpaar Italienisch-Südtiroler Deutsch zu testen. Südtiroler Deutsch ist an das italienische Rechtssystem gebunden. Im Laufe der Jahre hat sich eine spezifische rechtlich-administrative Terminologie entwickelt, die die Übersetzung dieser Sprachvarietät durch MÜ-Systeme besonders schwierig gestaltet. Die gewählte Domäne ist der Arbeitsschutz, ein Teilbereich des Rechts mit sehr besonderen terminologischen Merkmalen. Da der Arbeitsschutz in Südtirol ein zentrales Thema ist und jährlich zu diesem Thema Schulungen für Mitarbeiter verfasst und übersetzt werden, ist die vorliegende Masterarbeit ein wertvoller Versuch, den Übersetzungsprozess in dieser Domäne zu erleichtern und zu beschleunigen. Um ein vortrainiertes NMÜ-modell zu einer spezifischen Domäne zu adaptieren, wurden zwei parallele Korpora, bestehend aus etwa 25.000 Übersetzungseinheiten, vom Südtiroler Forschungszentrum *Eurac Research* zur Verfügung gestellt. Die Daten wurden bereinigt, gefiltert und in ModernMT eingefügt. Die Übersetzungsqualität von drei verschiedenen Testsätzen wurde automatisch bewertet und 79 domänenspezifische Übersetzungseinheiten wurden manuell analysiert, um die Genauigkeit und Fehler im Sprachfluss zu kommentieren. Die Ergebnisse zeigen eine überraschende Steigerung von 6,4 BLEU-Punkten und 3,5 METEOR-Punkten, sowie eine deutliche Verbesserung der terminologischen Genauigkeit. Das Vorhandensein zahlreicher terminologischer Fehler und Fehlübersetzungen zeigt jedoch, dass noch Verbesserungspotential besteht.