

Alma Mater Studiorum Università di Bologna
DIPARTIMENTO INTERPRETAZIONE E TRADUZIONE
Corso di Laurea magistrale in Specialized Translation (classe LM-94)

TESI DI LAUREA
in Technology for Translation

**Neural machine translation adaptation and automatic terminology evaluation:
a case study on Italian and South Tyrolean German legal texts**

CANDIDATO

Antonio Giovanni Contarino

RELATORE

Adriano Ferraresi

CORRELATORE

Eva Berta Maria Wiesmann

*Anno Accademico 2020/2021
Secondo Appello*

TABLE OF CONTENTS

INTRODUCTION	1
1. THE AUTONOMOUS PROVINCE OF BOLZANO / BOZEN – SOUTH TYROL	3
1.1 INTRODUCTION.....	3
1.2 LANGUAGE GROUPS IN SOUTH TYROL	3
1.3 HISTORICAL OVERVIEW.....	5
1.4 LEGAL TERMINOLOGY IN SOUTH TYROL.....	8
1.4.1 <i>The development of the German legal terminology in South Tyrol</i>	<i>10</i>
1.4.1.1 Early developments.....	10
1.4.1.2 The Terminology Commission	11
1.4.1.3 From standardisation to harmonisation.....	15
1.5 TRANSLATION IN THE SOUTH TYROLEAN PUBLIC ADMINISTRATION	16
1.5.1 <i>Translation policy.....</i>	<i>16</i>
1.5.2 <i>Translation technology and resources in South Tyrol</i>	<i>17</i>
1.5.2.1 Corpora	19
1.5.2.2 Translation memories	20
1.5.2.3 Terminology	21
1.5.2.4 Machine translation	22
1.6 SUMMING UP.....	23
2. MACHINE TRANSLATION	24
2.1 INTRODUCTION.....	24
2.2 HISTORICAL OVERVIEW.....	25
2.3 MT ARCHITECTURES	27
2.3.1 <i>Rule-based approaches.....</i>	<i>27</i>
2.3.2 <i>Corpus-based approaches</i>	<i>28</i>
2.3.2.1 Example-based machine translation	29
2.3.2.2 Statistical machine translation	29
2.3.2.3 Neural machine translation.....	31
2.4 PARALLEL CORPORA FOR MT TRAINING	33
2.4.1 <i>Cleaning noise in parallel data</i>	<i>34</i>
2.5 MACHINE TRANSLATION FOR INSTITUTIONAL TRANSLATION	38
2.6 NMT AND TERMINOLOGY.....	40
2.7 EVALUATION OF MACHINE TRANSLATION QUALITY	42
2.8 MACHINE TRANSLATION OF LEGAL-ADMINISTRATIVE TEXTS.....	46
2.9 TERMINOLOGY EVALUATION.....	47
2.10 SUMMING UP.....	50
3. METHODS.....	52
3.1 INTRODUCTION.....	52
3.2 AIMS AND RESEARCH QUESTIONS	52
3.3 BUILDING THE LEXB PARALLEL CORPUS	54
3.3.1 <i>Data collection and pre-processing</i>	<i>55</i>
3.3.2 <i>Segmentation and sentence alignment</i>	<i>57</i>
3.3.3 <i>Corpus cleaning and filtering.....</i>	<i>57</i>
3.3.3.1 Sentence-level cleaning.....	60
3.3.3.2 Corpus filtering.....	60
3.4 DOMAIN ADAPTATION	61
3.4.1 <i>Dataset splitting and near-duplicate processing</i>	<i>61</i>

3.4.2	<i>ModernMT</i>	63
3.5	OVERALL QUALITY EVALUATION.....	64
3.5.1	<i>Evaluation metrics</i>	64
3.5.2	<i>Statistical significance testing</i>	65
3.6	TERMINOLOGY EVALUATION.....	65
3.6.1	<i>Evaluation taxonomy</i>	67
3.6.2	<i>Data pre-processing</i>	71
3.6.3	<i>Term matching and test set creation</i>	72
3.6.4	<i>Evaluation and annotation</i>	74
3.6.5	<i>Advantages and limitations</i>	75
3.7	SUMMING UP.....	77
4.	RESULTS	78
4.1	INTRODUCTION.....	78
4.2	MT QUALITY EVALUATION	78
4.3	AUTOMATIC EVALUATION OF TERMINOLOGY TRANSLATION.....	80
4.3.1	<i>ModernMT baseline system</i>	80
4.3.2	<i>ModernMT adapted system</i>	82
4.4	DISCUSSION.....	89
	CONCLUSIONS	91
	REFERENCES	95
	APPENDIX A - CORPUS CLEANING AND FILTERING	122
	ACKNOWLEDGEMENTS	128
	ABSTRACT	129

INTRODUCTION

Following the implementation of South Tyrol's Statute of Autonomy, the public administrations of the Autonomous Province of Bozen/Bolzano are legally bound to use Italian and German as official languages and to publish laws and administrative acts in bilingual form. This results in a strong demand for translation of legal-administrative texts, usually from Italian into German, which could be satisfied, to some extent, by integrating machine translation (MT) in the institutional translation workflow. In this setting, local South Tyrolean legal-administrative terminology, which exhibits peculiar features with respect to other German-speaking countries, is of central importance in institutional translation. Previous studies have shown that legal terminology is the main type of error when machine-translating Italian legal-administrative texts into South Tyrolean German (Heiss and Soffritti 2018; Wiesmann 2019; De Camillis 2021).

The present work is part of a pilot project by the Institute of Applied Linguistics at Eurac Research and was partially carried out during a 300-hour internship in June and July 2021. The pilot project (MT@BZ)¹ consisted in a preliminary study on MT at South Tyrolean institutions. Despite the high translation demands of legal-administrative texts and the specific terminological needs related to the features of the local variety of German, the South Tyrolean administrations have not yet implemented any kind of MT system in their translation workflows. In this context, an MT system adapted to the local language and terminology could help to optimise local institutional translation processes and improve language accessibility. More specifically, the aims of the MT@BZ project include collecting local bilingual language data and carrying out domain adaptation of a selected MT system.

Within the MT@BZ project, the present work describes the pipeline that was implemented to collect, align and clean South Tyrolean bilingual data and to adapt an MT system (ModernMT), as well as to automatically evaluate a) the overall quality

¹ <https://www.eurac.edu/en/institutes-centers/institute-for-applied-linguistics/projects/mtbz> (last accessed 16/09/2021).

improvement of the engine, and b) the accuracy of legal terminology translation, thanks to a fine-grained taxonomy developed specifically for this task.

Chapter 1 is devoted to the setting of the study (the Autonomous Province of Bolzano/Bozen – Alto Adige) and provides a historical overview of the province and a review of the local linguistic, terminological and translation policies.

Chapter 2 is focused on machine translation and outlines the historical development of MT and the main MT architectures, with particular emphasis on neural machine translation. Moreover, the Chapter covers MT domain adaptation, data filtering for MT training, MT evaluation and MT use in institutional translation. Finally, related work on MT of legal-administrative texts and on terminology evaluation is reviewed.

Chapter 3 outlines the methodology adopted in the present dissertation project. The Chapter presents the aims and research questions of the study and provides details about the methods applied for parallel corpus building, MT adaptation, MT overall quality evaluation and automatic terminology evaluation.

Finally, Chapter 4 presents and discusses the results of the research, including results of the comparative quality evaluation between the generic and adapted MT system with regards to overall MT quality and the automatic evaluation of legal terminology accuracy in the MT outputs.

CHAPTER 1

1. THE AUTONOMOUS PROVINCE OF BOLZANO / BOZEN – SOUTH TYROL

1.1 Introduction

The present Chapter will be devoted to describing the setting of the study, the Autonomous Province of Bolzano/Bozen – South Tyrol, with a particular focus on the evolution of language policies, the peculiar terminology features of South Tyrolean German and today's institutional translation workflow in the South Tyrolean public administrations.

Firstly, details about the three language groups coexisting in South Tyrol will be provided, as well as a brief overview on the main historical events of the last century, which shaped today's political and linguistic situation of the province. Afterwards, the peculiar features of legal language and terminology in the South Tyrolean German language will be presented, along with a description of its development over the last century. Finally, the current translation policies adopted in the South Tyrolean public administration will be reviewed, based on the detailed studies recently carried out by Sandrini (2019) and De Camillis (2021). Particular focus will be put on translation technologies and the gaps in today's institutional translation workflow of South Tyrolean's public administration in terms of language resources and implemented technologies.

1.2 Language groups in South Tyrol

South Tyrol² is a trilingual province in Northern Italy with about 533,000 inhabitants (ASTAT 2020: 9). The majority of the population are native German speakers

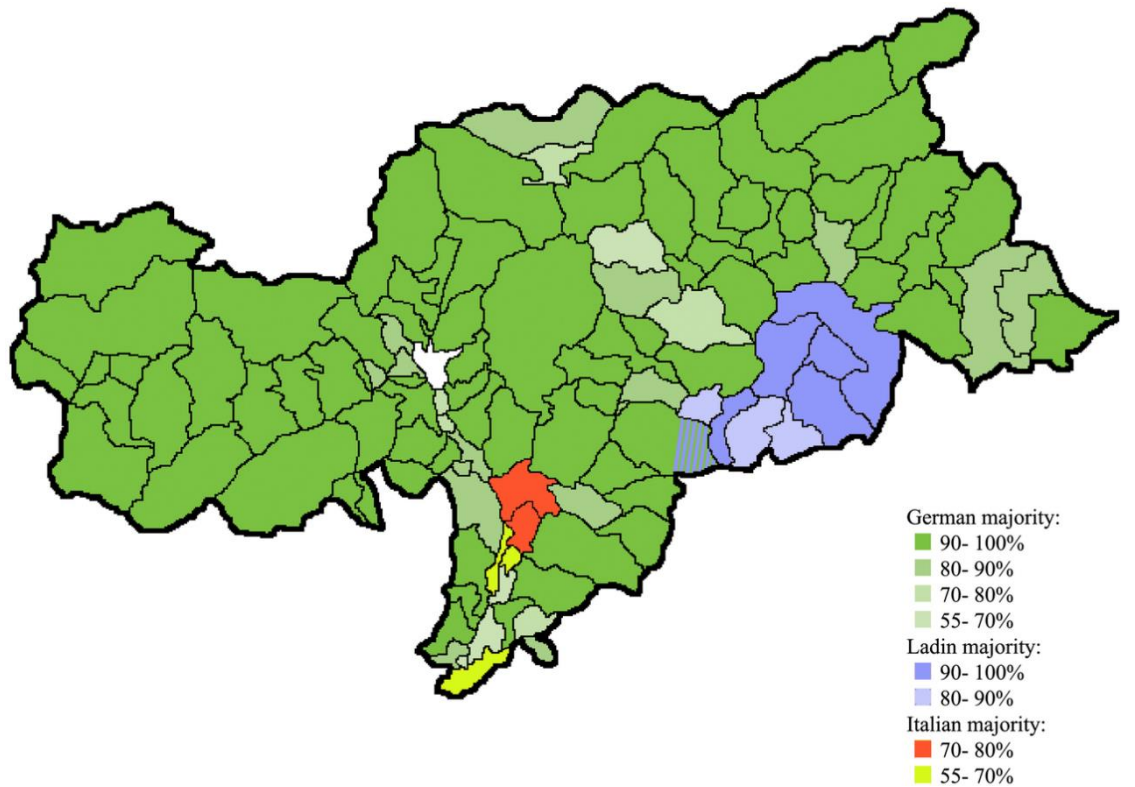
² The official Italian denomination is *Provincia Autonoma di Bolzano-Alto Adige*.

(69.64%), 25.84% are Italian speakers and 4.5% are speakers of Ladin³ (ASTAT 2020: 15). A strong concentration of Italian speakers is recorded mainly in the urban centres, while the peripheral and rural areas are almost completely German monolingual. The German language group⁴ is prevalent in 103 municipalities, whereas the Italian-speaking group holds the majority in only five municipalities (Bolzano, Laives, Salorno, Bronzolo, Vadena). In 26 municipalities the percentage of German native speakers is higher than 98% (ASTAT 2012: 10; see Fig. 1). German is the co-official language in the entire territory of South Tyrol. Legislative texts and other official documents are published in bilingual parallel version, but the only legally authoritative version is the Italian text. German-speaking citizens of the province of Bozen/Bolzano are entitled to use their own language when dealing with courts and public administration bodies and offices (Presidential Decree no. 670/1972, arts. 99-100). Employees in the public administration have to be bilingual and must undertake a language examination in their non-native language (Presidential Decree no. 752/1976). Schools of all levels are taught in Italian or German and second language courses are compulsory (Presidential Decree no. 670/1972, art. 19). The official denominations of place names are bilingual (Presidential Decree no. 670/1972, art. 8).

Ladin is co-official language in the valleys of Gardena and Badia, each with a local variety of Ladin. In the municipalities of these valleys, the Ladin language can be used when dealing with public administration offices. At the provincial level, relevant legislative and administrative texts are drafted or translated in one of the two variants of the Badia and Gardena valleys, weighing an equal presence of both variants (Resolution of the Provincial Government no. 210/2003). Place names are trilingual (Videsott *et al.* 2020: 50). Schools are taught both in German and Italian, each for the same number of hours, whereas Ladin is taught a few hours a week (Presidential Decree no. 670/1972, art. 19; Videsott *et al.* 2020: 37).

³ Ladin is a romance language spoken by about 32,500 speakers in five valleys (Badia, Gardena, Fassa, Livinallongo, Ampezzo) in the provinces of Bolzano, Trento and Belluno (Videsott *et al.* 2020: 3, 36). About 20,000 speakers of Ladin live in South Tyrol (ASTAT 2012: 4).

⁴ In multilingual regions and countries (like Switzerland and South Tyrol), a “language group” is defined as the group of speakers of a given language.



White: Merano (It. 49.06%, Ger. 50.47%, Lad. 0.47%)

Figure 1: Language groups in South Tyrol – Census 2011 (ASTAT 2012)

1.3 Historical overview

German is considered a minority language in Italy, but, for historical reasons, it is the most used language at the local level in South Tyrol. For more than a millennium, South Tyrol has been a borderland region inhabited mainly by a German-speaking population. In 1363, the previously independent county of Tyrol came under the rule of the Habsburg dynasty. From then on, the region was part of Austria (it was assigned to other states for during the Napoleonic era, but only for a few years) until the end of the First World War in 1919 (Peterlini 2000: 33–34, Gruber 2000: 13).

After the end of the First World War and the defeat of Austria, the Tyrol historical region was divided without any safeguard clause for South Tyrol (Stocker 2006: 16). The separation, which was decreed by the peace treaty of Saint Germain on 20 September 1919, did not take into account the principle of self-determination of peoples as laid down in Wilson's Fourteen Points. Moreover, the division did not take place “along clearly recognisable lines of nationality” (Wilson 1918), which could only be the linguistic boundary of the Salurn Gorge, but along the watershed line at the Brenner

Pass. Thus, the territories annexed to Italy included not only the southernmost part of Tyrol (today's *Trentino*) but also the predominantly German-speaking area of Tyrol called *Südtirol* (Alcock 2001: 1–2; Gruber 2000: 15).

The peace treaty did not provide any protection for the German language group in the annexed territories and South Tyrol was aggregated with Trentino to form a single province called "Venezia Tridentina" (Alcock 2001: 2). In those years, a first attempt of Italianisation was undertaken, although King Vittorio Emanuele III had declared that the Italian state should respect the local autonomy and traditions (Peterlini 2000: 67–68). Several South Tyrolean employees of the public administration were dismissed, a number of German-speaking schools were closed and the so-called "small option"⁵ took place: thirty thousand people were not automatically granted Italian citizenship and ten thousand of them were subsequently denied it; this led to a first exodus of the German-speaking population (Stocker 2006: 18–19).

In the period between 1922 and the First World War, the Fascist policy was to assimilate South Tyroleans and to forcibly "Italianise" South Tyrol (Alcock 2001: 2; Stocker 2006: 20). Italian was made the only official language. The German language was banned from public life and those who did not speak and write Italian were dismissed from their posts. German newspapers were closed down. Court cases had to be heard in Italian only. In schools, the only teaching language allowed was Italian, and South Tyrolean children were secretly taught German in so-called "catacomb" schools. All South Tyrolean signs, inscriptions and place names were Italianised. Forced Italianisation measures included a strong inflow of workers and employees from other regions of Italy (Alcock 2001: 2–3; Gruber 2000: 33–35; Peterlini 2000: 69–70; Stocker 2006: 20–22).

In 1939, with the Hitler-Mussolini Options Agreement, South Tyroleans were given the option to either leave their homeland to be resettled in the German *Reich* or remain in South Tyrol and accept complete assimilation. The majority of the South Tyroleans (86%, i.e., more than 200,000 people) opted for German citizenship and

⁵ "*Kleine Option*", a reference to the 1939 Hitler-Mussolini Options Agreement (see below in this Section).

therefore chose to emigrate, while those who remained were considered traitors and despised. However, probably also due to the outbreak of World War II, only 75,000 South Tyroleans actually emigrated, and about 20,000 of them returned to South Tyrol after the end of the war (Alcock 2001: 3-4; Stocker 2006: 29–35).

In 1946, after the end of World War II, Italy and Austria signed an agreement (called Degasperi-Gruber Agreement, or Paris Agreement) which laid the foundations for the South Tyrolean autonomy. The Paris Agreement granted the German-speaking population of the area the right to autonomy and to preserve its ethnic and cultural character. It ensured the parification of the German and Italian languages in public offices and official documents, as well as complete equality between German-speaking and Italian-speaking people in Trentino and South Tyrol, recognising, among others, the right to be taught in one's mother tongue (Alcock 2001: 4–6; Gruber 2000: 103).

Two years later, the First Statute of Autonomy was promulgated in order to implement the provisions of the Paris Agreement. Despite granting some important language rights, the First Statute of Autonomy left the South Tyrolean minority population unsatisfied. The autonomy was bestowed on both the South Tyrolean territory and the mainly Italian-speaking province of Trento, making the German-speaking population the minority within the regional parliament. Moreover, using the German language in public life remained a mere possibility and Italian remained the only official and legally binding language (Alcock 2001: 6–7; Chiocchetti *et al.* 2017: 255–256; Gruber 2000: 107–109). In the following two decades, the unrest among the German-speaking population continued to rise, calling for a separation from Trento and Italy. The tension escalated into violence and bomb attacks by extremist groups starting from 1956 (Alcock 2001: 8; Peterlini 2000: 91–93; Stocker 2006: 55–58).

The situation changed radically with the promulgation of the New Statute of Autonomy (1972), which conferred autonomy to each of the two provinces of Bolzano and Trento and transferred previously regional powers to the provincial parliaments. Even though Italian remained the official language of the State, the New Statute of Autonomy recognised equal status to the German and Italian languages at the local level and granted the German-speaking minority the right to use their language when communicating with the public administration and in courts (Presidential Decree no. 670/1972, arts. 99-100). The ethnic proportion, i.e., the distribution of public posts

according to the proportions of speakers of the language groups living in South Tyrol,⁶ was introduced in public employment in order to protect minority language groups. Most official legal and administrative documents produced by the public administrations had to be published in both Italian and German. Moreover, to ensure bilingualism in the public administration, a language exam was later made mandatory for access to public employment (Alcock 2001: 10–14, 17; Peterlini 2000: 112ff.). Since the New Statute of Autonomy, the local government has obtained growing autonomy and primary or secondary competencies and, today, it legislates on almost all matters of public life (Alcock 2001: 11–18; Peterlini 2011: 306).

1.4 Legal terminology in South Tyrol

The South Tyrolean public administrations are by definition *translating institutions*,⁷ since they are “regulatory organizational systems [...] that operate in a multilingual environment” that “employ translation in performing their governing function” (Koskinen 2014: 3). Legal terminology is a key aspect in this context, since institutional translation mainly deals with legal-administrative texts (Koskinen 2011: 57; De Camillis 2021: 31).

As Ralli (2009) explains and exemplifies, legal terminology is very heterogeneous. It may include terms and formulaic expressions that are considered purely legal (such as “synallagma”), as well as terms borrowed from ordinary language, which undergo a semantic redetermination and acquire legal value if placed in a legal context. Moreover, as legal language is the language of law, which “develops through a complex network of legal branches that encompass virtually every aspect of life” (Prieto Ramos 2021: 176), legal terminology is interdisciplinary, i.e., it also includes terms

⁶ Ital. *proporzionale etnica*.

⁷ Koskinen defines *institutional translation* as “those cases when an official body [...] uses translation as a means of ‘speaking’ to a particular audience” (2008: 22). Koskinen later restricts the concept of institutional translation to “those concrete institutions that directly serve the societies’ control and governance functions” (2011: 57). The main examples of institutional translations are “official documents of government agencies and local authorities of bilingual or multilingual countries; translating in the European Union, the United Nations and other international or supranational organizations, and international courts of law” (*ibid.*).

borrowed from other disciplines (e.g., terms from medicine in family law, like “gametes” and “assisted reproduction”; terms from engineering, insurance and transport in road law; etc.) (Ralli 2009; Prieto Ramos 2021: 176). The constant interaction of the legal field with other fields of knowledge results in terms being sometimes re-defined at the normative level, even though they are already defined in their specialized domain, causing what Soffritti (2002: 60 in Ralli 2009) calls a “double linguistic specialisation” (*doppelte Fachsprachlichkeit*) of terms.

Legal language is the expression of a specific culture and is always bound to the legal system it pertains to (Sandrini 1996: 138; Wiesmann 2004: 19). Each legal system is characterized by its own conceptual structures and develops its own legal terminology to express concepts and to pursue its regulatory objectives (Sandrini 1996: 138). Therefore, full terminological equivalence between terms pertaining to different legal systems is rare (Sandrini 1996: 138; Wiesmann 2004: 233; Chiocchetti and Ralli 2013: 11). As a consequence, different countries and legal systems that use the same official language (for instance, German-speaking countries and regions⁸) may be characterised by partially or totally divergent legal concepts and terms, as well as by many unique legal institutions (Chiocchetti and Ralli 2016: 103). This is the case of South Tyrol, where both official languages, Italian and German, are strictly bound to the Italian legal system. However, before the annexation of South Tyrol to Italy, the German language had never been developed to express concepts pertaining to the Italian legal system, as there had not been a direct connection between German and the Italian law before (Coluccia 2000: 381; in Chiocchetti 2019b: 177). Therefore, also due to its historical, legal and linguistic background and to the approaches adopted towards legal terminology in the last 50 years (see Section 1.4.1), the German language in South Tyrol features a number of distinctive peculiarities and represents a unique case within

⁸ German is (considered) a pluricentric language, i.e., it is recognised as an official language in several countries and regions. In particular, German is an official national language in Germany, Austria and Lichtenstein, an official regional language in eastern Belgium and South Tyrol, and a co-official language in Switzerland and Luxembourg. As a pluricentric language, German comprises three national full centres (Germany, Austria and German-speaking Switzerland), with a codified language standard, and four half centres (Lichtenstein, Luxembourg, eastern Belgium and South Tyrol) (Ammon 1995: 96; Ammon *et al.* 2004: XXXII; Thüne *et al.* 2011: 48–49)

the German-speaking linguistic area (Chiocchetti *et al.* 2013a: 262; Chiocchetti and Ralli 2016: 103).

1.4.1 The development of the German legal terminology in South Tyrol

1.4.1.1 Early developments

In the last century, the German legal-administrative language in South Tyrol experienced different stages of evolution, which are strictly connected with the historical and political background of the region (see 2.1.3). Until the 1960s and early 1970s, the German language remained mostly absent from legal and administrative documents, as the use of German in those settings was not fully granted until the promulgation of the New Statute of Autonomy (1972). Therefore, since legal terminology in German language was rarely used in that period, no relevant linguistic development took place at that stage (Zanon 2001: 168; Chiocchetti *et al.* 2017: 256; Chiocchetti and Ralli 2016: 104).

From the years before the New Statute of Autonomy until the 70s and 80s, several public and private organisations carried out translations of legal and administrative texts from Italian into German, including the main legal codes and application forms used in the public administration and judiciary system. This translation activity, however, was carried out over time by different translators, without any coordination or attempt at systematisation with regards to terminology. Although leading to a *de facto* standardisation of German equivalents of Italian legal terms (Sandrini 1998: 411), this translation activity also caused growing terminological variation, with the parallel use and co-occurrence of several different, often incorrect designations for the same concept (Chiocchetti and Ralli 2016: 105; Zanon 2008: 54; Chiocchetti 2019a: 105). At that stage, characterized by what Chiocchetti calls an “uncoordinated and unplanned ‘laissez-faire’ approach to legal and administrative terminology” (2019a: 106), existing legal terminology in South Tyrolean German consisted of:

- a) Terms of Austro-Hungarian origin, which referred to concepts that were unknown to the Italian legal system, in particular terminology related to land

registry (*Grundbuch*)⁹ (Zanon 2001: 177; Chiocchetti 2019b: 177; Chiocchetti and Ralli 2016: 104).

- b) Terms generated spontaneously from everyday use of the population, which used to express concepts pertaining to the Italian legal system that were formerly unknown to the German-speaking inhabitants. These terms were mainly adapted loan words (*Quästur* from the Italian *questura*, “police headquarters”) or loan translations (*Autobüchlein* from the Italian *libretto di circolazione*, “vehicle registration certificate”). Some of these terms have remained in use for many years and it is now hard to propose alternatives that would be accepted and adopted systematically by the population (Chiocchetti 2019b: 177; Chiocchetti 2019a: 106; Chiocchetti and Ralli 2016: 104).
- c) Terms contained in the translations into German of Italian legal texts carried out by public and private organisations (see above in this Section). As pointed out earlier in this section, since translations were made by many translators, at different times and with no systematic coordination or terminology planning, these terms often included several variants, which not rarely turned out to be incorrect (Zanon 2008:54). For example, the term *assessore provinciale* was translated over time both as *Provinzialassessor*, *Landesassessor* and *Landesrat* (Chiocchetti 2019a: 106).

1.4.1.2 The Terminology Commission

With the promulgation of the New Statute of Autonomy in 1972, German achieved fully equal status to the Italian language also in court and when communicating with the public administration (see Section 1.3). However, bilingual communication in the legal domain requires that all legal and specialised terms have adequate equivalents in the second language and are used consistently (Zanon 2001: 176). As a consequence of the approach adopted during the early developments of legal terminology in South Tyrol (see Section 1.4.1.1), instead,

⁹ The land registry (*Grundbuch*) system is different from the Italian cadaster system (*Kataster*), but both systems are today present in formerly Austro-Hungarian areas of Italy (Chiocchetti 2019b: 177).

“[t]he available material lacked coherence, was largely incomplete, was often very much based on the Italian terminology (loan words and loan translations), was not always correct from a legal and linguistic point of view, but sometimes was already consolidated in daily use.”

(Chiocchetti 2019b: 178)

This situation caused what Chiocchetti *et al.* call a “terminological emergency” (2013c: 22). In order to ensure the use of the German language in court and in the public administration “nicht nur *de jure*, sondern auch *de facto*”¹⁰ (Chiocchetti and Ralli 2016: 104), the need for the development of a unified and harmonised terminology in German language, “welche die italienischen Rechtsinstitute präzise und gleichzeitig wahrheitsgetreu abbildet”¹¹ (Woelk 2000: 213), became evident. The problem, in other words, was to effectively express the Italian legal culture in German (Chiocchetti 2019a: 106; Chiocchetti and Ralli 2016: 104).

To address this situation of “terminological emergency”, in 1988 a Terminology Commission of three German-speaking and three Italian-speaking experts was established (Presidential Decree no. 574/1988). The task of the Commission, which started its works in 1991, was “to retrieve, check, approve, and update the legal, administrative, and technical terminology necessary for the local administrative and judiciary system”, with the aim of “establishing official German equivalents for the existing Italian legal terminology” (Chiocchetti 2019a: 106); Presidential Decree no. 574/1988, art. 6).

The preliminary work for standardisation was carried out by terminologists and lawyers of the European Academy of Bolzano (EURAC) and consisted in the legal comparative analysis of Italian concepts with German-speaking legal systems, to verify the existence of equivalent concepts and respective designations. Existing and established South Tyrolean specialised terminology was taken into account as well. If existing German equivalent terms could not be identified, translation proposals were

¹⁰ “not only *de jure*, but also *de facto*” (my translation).

¹¹ “[...] that depicts the Italian legal institutions precisely and at the same time truthfully” (my translation).

provided.¹² After undergoing a revision process carried out by domain experts, the proposed entries were then submitted to the Terminology Commission for approval. The decisions of the Terminology Commission were forwarded to the local government and to the *Commissario del Governo*, who could propose changes within six months (Alber and Palermo 2012: 297–298; Chiocchetti *et al.* 2017: 259–261; Chiocchetti 2019a: 106–107; Chiocchetti and Stanizzi 2010: 2–3; Mayer 2000: 297–298; Zanon 2008: 55–57). Finally, terms were standardised and published as one-to-one correspondents and became legally binding in all texts written by public authorities: this represented a clear example of a “prescriptive approach to terminology planning and management, since it meant indicating even more than just a ‘preferred usage’” (Chiocchetti 2019a: 106).

In order to establish German equivalent terms, the Commission (and, during preliminary terminology work, terminologists at Eurac) needed to consider terminology from other German-speaking legal systems (Mayer 2000: 299–300; Chiocchetti 2019a: 107). Although they could not be considered “as real models for transposing Italian terminology into German” (Alber and Palermo 2012: 299), existing terms from the German, Austrian and Swiss legal systems could be used as reference points and adopted as South Tyrolean equivalents, as long as “substantial conceptual equivalence between the Italian and foreign legal concepts could be established” (Chiocchetti 2019a: 107). When no adequate functional equivalence was found in other legal systems, the terminological gap could be filled by either maintaining the term in the original language (as in the case of *Carabinieri*) or by creating a neologism, usually loan translations of the Italian term or German paraphrases of the respective concept (Ralli 2009; Chiocchetti 2019a: 107). Coining new terms from scratch, however, could have a foreignizing effect and bring about an excessively strong regionalization (Sandrini 1998: 408). When proposing new terminology, moreover, the Commission was often faced with the dilemma of either choosing terms that were as precise as possible from a legal point of view (sometimes resulting in overlong paraphrases) or more accessible

¹² For a more in-depth overview on comparative legal terminology work carried out by terminologists at Eurac Research, see (Chiocchetti *et al.* 2013b; Chiocchetti *et al.* 2013c; Chiocchetti *et al.* 2019).

but imprecise solutions (Zanon 2008: 59). While seeking a balance between the legal correctness of terms and their accessibility, however, the former often prevailed (Ralli and Stanizzi 2018: 178).

Between 1991 and 2012, the Terminology Commission standardised approximately 7,400 terminological equivalents, which were published in twelve separate term lists. Moreover, bilingual terminology contained in the parallel editions of the main Italian legal codes translated in the preceding decades (see Section 2.1.3.1.1) was “batch standardised”, bringing the total number of standardised terms to approximately 15,000 to 20,000 terms (Chiocchetti 2019c: 10–11; Chiocchetti *et al.* 2017: 260, 265).

The standardisation activity of the Terminology Commission, which was discontinued in 2012, had both positive and negative effects on the development of legal terminology in South Tyrol. On the one hand, it provided legal drafters and translators with a consistent corpus of reference terms, therefore facilitating communication at the administrative level and helping to improve the comprehensibility and accuracy of legal translations in the public sector. Moreover, standardisation limited the proliferation of synonyms in favour of unambiguous and binding translations (Ralli and Stanizzi 2018: 176–177; Chiocchetti 2019a: 107). On the other hand, however, the standardisation process was slow and complex, making it difficult to keep up with the constant conceptual and linguistic evolution of law and to guarantee a continuously updated terminology (Ralli and Stanizzi 2018: 176–177; Chiocchetti 2019a: 107). Also due to the reduced number of Commission members, who only worked on a part-time basis, the amount of bilingual standardised terms was deemed unsatisfactory, with several domains which remained largely uncovered (Zanon 2008:58). Moreover, standardised terminology was (and still is) not systematically adopted by users and even South Tyrolean public institutions, which in some cases prefer using neologisms or outdated terms in place of the standardised term (Alber-Palermo 2012:301, Zanon 2008:58). As an example of this tendency, *Quästur*, an adapted loan word from the Italian *questura*, is still widely used by the population and insiders, and can be found in place of the official standardised translation *Polizeidirektion* even on the signpost outside the Bolzano police headquarters (Chiocchetti *et al.* 2017: 262; Chiocchetti 2019a: 107).

1.4.1.3 From standardisation to harmonisation

After the Terminology Commission ceased its activity in 2012, the need to reform the standardisation process became clear. This was mainly due to the necessity of overcoming some of the limitations of the terminology standardisation work carried out by the Commission until then (see Section 1.4.1.3). The slow and complex standardisation process had “by far not met all the terminological needs of the German-speaking community in South Tyrol” (Chiocchetti 2019a: 108) and many standardised terms had already become obsolete and needed to be updated. Moreover, the issue of terminology introduced by newly issued legislation had to be addressed as well, since the South Tyrolean context highlights not only the need for a terminology base covering the various areas of law, but also the practical need for a correct, unambiguous, constantly updated and easily accessible terminology (Ralli and Stanizzi 2018: 179).

Against this background, in 2015, terminologists at Eurac Research and the Office for Language Issues¹³ signed a cooperation and data exchange agreement, with the aim of “moving from prescriptive to descriptive and translation-oriented terminology work”, by targeting “high-quality terminology work to current needs” in order to “update and integrate existing terminologies more quickly” (Chiocchetti *et al.* 2017: 267). In light of that cooperation, terminology work today follows a descriptive approach and is based on the systematic processing of the terminology contained in normative texts translated and revised by the Office for Language Issues. Moreover, ad hoc terminology work is carried out on single terms or small groups of terms on a needs basis: this happens, for example, when new legislation is issued by the central government and the corresponding South Tyrolean German terms need to be defined quickly, in order to avoid the creation of many different translations in the South Tyrolean context (Chiocchetti 2019a: 108; Ralli and Stanizzi 2018: 181). As a result of this terminology activity, which is still carried out applying the method of legal

¹³ The Office for Language Issues (*Ufficio Questioni Linguistiche*) is the main language advisor to the province and is the only office within the provincial administration which formally carries out translation tasks. It is responsible for the linguistic revision of normative texts and other documents of the provincial administration, as well as for the collection and management of terminology specific to the areas of competence of the provincial administration. The website of the Office can be visited on https://www.provincia.bz.it/it/contatti.asp?orga_orgaid=472.

comparison to other German-speaking legal systems, recommended terms for South Tyrol are flagged with the label "*in Südtirol empfohlen*" (recommended use in South Tyrol) (Ralli and Stanizzi 2018: 182). The respective terminology entries are published in the online Information System for Legal Terminology *bistro*¹⁴ (Ralli and Andreatta 2018). Flagged terms are not legally binding, but only recommended for use in South Tyrol, and terminology work carried out today in South Tyrol has therefore moved from standardisation to terminological harmonisation¹⁵ (Chiocchetti 2019c: 14; De Camillis 2021: 153–154).

1.5 Translation in the South Tyrolean public administration

1.5.1 Translation policy

In South Tyrol, by law, a number of legal-administrative documents must be published in bilingual version. In particular, Presidential Decree 574/1988 prescribes the joint use of Italian and German when issuing: a) acts intended for the general public that are to be published (e.g., laws); b) individual acts intended for public use (such as identification documents, concession acts, etc.); c) acts intended for multiple public administration offices. Bilingual texts must be displayed side by side and must have the same typographical layout (art. 4, par. 4). However, the legally binding version is always the Italian text (Presidential Decree 670/1972).

The bilingual compilation of the acts to be published is carried out by the bodies and offices responsible for publication (Presidential Decree 574/1988, art. 5). Only normative texts (acts intended for the general public, see above in the same Section) have to undergo a linguistic revision, carried out by the Office for Language Issues, within their drafting process (De Camillis 2017). The translation of other institutional documents, e.g., administrative acts, non-binding informative documents (e.g., reports), or other technical texts are, on the contrary, entirely managed within individual departments of the administration. As there is no normative prescription regarding the

¹⁴ <http://bistro.eurac.edu/>. See Section 1.5.2.3.

¹⁵ Unlike standardisation, harmonisation does not entail the approval of terminology by a standardising body (ISO 860:2007; ISO 10241-2:2012).

translation policy of this type of documents, each office or department can decide whether to draft them in two languages in parallel, translate them internally or outsource the translation (De Camillis 2021: 96–97).

In the South Tyrolean public administration, therefore, there is no central unit in charge of translation. The Office for Language Issues, which is the only kind of central translation unit for the local administration (Sandrini 2019: 303), only occasionally deals with translations, whereas it is mainly responsible for linguistic revision in the process of drafting normative texts (De Camillis 2017). Furthermore, individual departments do not have any translation office or other language service, nor do they have staff employed as translators (De Camillis 2021: 97). The majority of translations at the administrative level is therefore carried out by employees, as translation is considered one of the general tasks of all public employees (Collective Agreement 08/03/2006, art. 2). Most of these employees-translators are untrained: 67% of the employees of the public administration translate as non-professional translators, and only 6,5% of employees-translators with a university degree has a degree in a language-related field (De Camillis 2021: 204; 212). As we can see, in the South Tyrolean public administration, translation competence is associated exclusively with linguistic competence (Sandrini 2019: 344).

In general, an explicit translation policy in the South Tyrolean administration is nearly absent and still has much room for improvement (Sandrini 2019: 343). The main causes of this situation have been identified in the overlap between translation competence and linguistic competence and the consequent lack of professionalisation of the role of translators, in the lack of a central translation unit and in the insufficient coordination and cooperation between existing translation offices within different administration departments (Sandrini 2019: 343–375). Moreover, as we will see in the following Section, in the translation processes taking place in the South Tyrolean administration there are also important shortcomings with regards to the implementation of translation technologies.

1.5.2 Translation technology and resources in South Tyrol

Translation technology, i.e., the different types of technology used in human translation, machine translation and computer-aided translation, has brought radical changes to all

aspects of the contemporary world of translation (Chan 2015: xxvii–xxviii), bringing about what has been called a “technological turn” in translation practice and translation studies (Cronin 2010).

The benefits of applying translation technologies are mainly related to an increased quality and productivity, which includes time saving, cost reduction, reuse of resources, task-sharing and cooperation between translators, as well as terminology harmonisation (Sandrini 2019: 111–116; De Camillis 2021: 39). In a context characterised by the presence of minority languages (like South Tyrol), however, the advantages of translation technologies are not limited to higher efficiency and productivity. As Sandrini points out (2019: 111–112), in addition to the general goal of maintaining identity through the use and application of the language, the use of appropriate translation technology also allows for the digital storage of the minority language and, consequently, the availability of digital texts in both the minority and majority languages. Translation technologies can play a significant role in the development and growth of regional or minority languages, but they are essential for any institutional context, considering the high degree of textual standardisation (De Camillis 2021: 39).

In South Tyrol, however, the use of translation technologies is still severely lacking from many perspectives and represents the most defective aspect of the provincial administration's translation policy (Sandrini 2019: 376–378; De Camillis 2021: 91; 231). When compared to other multilingual regions with a comparable linguistic and political situation,¹⁶ South Tyrol turns out to be the less advanced in terms of implementation of translation technologies (De Camillis 2021: 159–162). This is mainly due to the lack of a centralised translation workflow with shared tools, resources and data, as well as to the lack of a coordinated translation data exchange policy, of defined guidelines and of a common strategy at the level of local administrations in relation to the application of translation technologies (Sandrini 2019: 363; 378). For the

¹⁶ De Camillis (2021) analysed the translation policies of Catalonia and the Basque Country and scored them based on the Translation Policy Metrics model (Sandrini 2019), comparing the results with the scores assigned to South Tyrol by Sandrini (2019). With regards to translation technology and resources, South Tyrol was assigned an overall score of 15/45, against the 22/45 score obtained by both Catalonia and the Basque Country (De Camillis 2021: 164)

same reasons, translation resources like corpora, translation memories and terminology databases are lacking or are not fully exploited by institutional translators (Sandrini 2019: 378–380; De Camillis 2021: 91, 161–162).

1.5.2.1 Corpora

Nowadays, the use of monolingual and parallel corpora¹⁷ is central not only in translation studies (Baker 1995) but also in translation practice (Zanettin 2002; Zanettin 2014: 188–190). Parallel corpora are mainly exploited by translators in the form of translation memories, a “specific type of dynamic parallel corpora” (Zanettin 2014: 179), but monolingual and parallel corpora can also be usefully queried using “corpus analysis software to find information about terms, phraseology and textual patterns in both source and target languages” (Zanettin 2014: 180), especially if texts are annotated with linguistic information (POS tags, lemmas). Moreover, the creation and exploitation of corpora is an important preliminary task for other translation technology-related applications, such as machine translation and terminology extraction (Sandrini 2019: 160).

In South Tyrol, local legislation is published online in the LexBrowser system and in the Official Bulletin. The LexBrowser¹⁸ database contains provincial normative texts, resolutions and constitutional legislation of interest to South Tyrol in the Italian, German, and sometimes Ladin version, as well as decisions of the Constitutional Court and the Regional Administrative Court, which rarely are translated into German. Single texts or articles can be retrieved by keywords, drafting year, article, number or text type. Texts, however, can only be consulted monolingually: a link to the translated version(s) exists, but it is not possible to display the original and the translated text side by side. The Official Bulletin is the exclusive means of institutional communication and legal

¹⁷ A corpus is defined as a “collection of (1) machine-readable (2) authentic texts [...] which is (3) sampled to be (4) representative of a particular language or language variety” (McEnery *et al.* 2006: 5). Monolingual corpora are limited to one language (McEnery and Hardie 2012: 18), whereas a parallel corpus “contains native language (L1) source texts and their (L2) translations (McEnery and Hardie 2012: 20).

¹⁸ <http://lexbrowser.provinz.bz.it/>.

publicity of regional laws and regulations, administrative acts and all acts of the Autonomous Region of Trentino-Alto Adige/Südtirol.¹⁹ All weekly publications, however, are issued in a bilingual PDF format only, making it difficult to quickly and effectively query them for translation purposes.

These text collections contain a large number of texts and represent a potentially useful resource for translators. However, aside from not being available in a translation memory format, which could be easily usable by translators within their translation workflow, the current configurations of the LexBrowser and the Official Bulletin do not make it possible to query texts effectively in their parallel format (Sandrini 2019: 378).

As also pointed out by Sandrini (2019: 392) and De Camillis (2021: 320–321), parallel text display, parallel concordance search and/or availability of texts in a standard translation memory format (TMX)²⁰ would be of huge usefulness for all translators of the public administration.

1.5.2.2 Translation memories

Today, translation memory²¹ systems are the most widely used translation technology application in the translation and localization industry (Reinke 2018: 55–56). Translation memories are the most important function of translation environment tools and constitute the basis of an efficient translation activity, as they allow translators to quickly retrieve previously translated sentences (Melby and Wright 2015: 364; Sandrini 2019: 379). The efficiency of exploiting translation memories becomes particularly evident when dealing with highly repetitive texts (Zanettin 2014: 189). Translation memories can be created either by adding new translation units while translating or by aligning existing translations and their original texts (Zanettin 2014: 194–195; Reinke 2018: 163). Although less sophisticated than annotated parallel corpora queried with

¹⁹ <http://www.regione.taa.it/burtaa/it/info.aspx>.

²⁰ The TMX (Translation Memory eXchange) format (LISA 2005), developed by the OSCAR group of the Localisation Industry Standards Association (LISA), is an XML markup formalism that allows any tool using translation memories to import and export databases between their own native formats and a common format (Melby and Wright 2015: 669).

²¹ A translation memory (TM) is a “database containing a collection of paired source language (SL)/target-language (TL) text units” (Melby and Wright 2015: 662).

concordancers, translation memories can also be used to generate parallel concordances of single words or expressions in the source text, by retrieving all translation units containing the desired word or expression (Zanettin 2002: 11; Zanettin 2014: 189).

As Sandrini (2019: 379) observed, translation offices of the South Tyrolean public administration seem reluctant to adopt translation memory systems. The use of translation memories is still very limited and sparse, and there is a lack of any coordinated data exchange among translators of different departments. Public administration could benefit extensively from the systematic and centrally managed use of translation memories. This would make it possible to reuse translations that have already been produced in the past in a quicker and more efficient way, especially since most of the texts translated by the translator-employees are highly repetitive and have a fixed structure (forms, resolutions, decrees, circulars and general communication) (De Camillis 2021: 219). Overall coherence, including terminological consistency, would also be enhanced (De Camillis 2021: 235). A first step in this direction could be the semi-automatic alignment of the bilingual texts available to the public administration to create a centralised translation memory (Sandrini 2019: 378).

1.5.2.3 Terminology

Terminology policies are of particular importance in the context of regional or minority languages, since a minority language can only be considered on an equal footing with the majority language if terms in legislation, public life and administration have adequate equivalents in the second language and are used consistently (Zanon 2001: 176; Sandrini 2019: 169). Terminology policy must therefore ensure that terminology work, terminology harmonisation and/or terminology standardisation are carried out and that the result of this work is openly available to translators and to the public (Sandrini 2019: 169; 251).

In South Tyrol, although terminology standardisation and harmonisation have been carried out for the last 30 years (see Sections 1.4.1.3 and 1.4.1.4), translation-oriented terminology management in the translation workflow of the public administration is still largely unsystematic, scattered and sometimes inexistent (Sandrini 2019: 380; De Camillis 2021: 91, 161–162, 232). There is virtually no coordination of terminology work or exchange of data among bilingual employees and translators from

different departments, and the connection to the official terminology work carried out by the Terminology Commission or Eurac Research is not systematic and coordinated (Sandrini 2019: 380). An exception is represented by the cooperation between the Office for Language Issues and the Institute for Applied Linguistic at Eurac Research (see Section 1.4.1.4), with the latter carrying out systematic terminology processing and legal comparison from recently issued legislation and publishing the results of terminology work in the *bistro* database (Ralli/Stanizzi 2018:179-186).

The *bistro*²² Information System for Legal Terminology was first developed in 2001 as a support tool for communication, writing and translations within a legal context. It is a concept-oriented termbase containing more than 11,000 fully-fledged terminological entries (Ralli/Andreatta 2018:9), each provided with definitions and contexts of use with respective sources, grammatical information, information about the term status (standardised, recommended, obsolete, etc.), information about the geographical use of German terms (South Tyrol, Germany, Austria, etc.), collocations, concept-level and term-level notes and cross-references to hypernym, hyponyms, co-hyponyms and related concepts.

1.5.2.4 Machine translation

Machine translation (MT) is undoubtedly one of today's most important translation technology tools (Sandrini 2019: 334). Thanks to the significant improvements made in the last years, especially after the paradigm shift towards neural machine translation (Bentivogli *et al.* 2016), MT is rapidly taking over the translation industry and is being adopted in several institutional and government settings as well (see Section 2.5).

As Sandrini (2019: 376–378) and De Camillis (2021: 91–159) observed, in the South Tyrolean public administration there seems to be a general opposition to MT: the use of MT systems is almost absent and, accordingly, no MT post-editing activity is carried out when translating official documents. Furthermore, there has not yet been any attempt to adapt an automatic translation system to the linguistic and terminological

²² <http://bistro.eurac.edu>.

needs of South Tyrol (by means of in-domain parallel corpora and/or terminology), despite the fact that public administration employees highlighted the need for such solution (Sandrini 2019: 363, 377; De Camillis 2021: 159, 233–234), since commercial MT systems don't effectively translate South Tyrolean legal terminology (Heiss and Soffritti 2018; Wiesmann 2019; De Camillis 2021: 291–300). Implementing MT and post editing within the translation workflow of South Tyrol's administration, in particular an MT system adapted to the local legal-administrative texts, would bring significant benefits in terms of productivity and, therefore, of reduced time and costs (Sandrini 2019: 192–193, 243), especially since MT of administrative and legal documents more readily produces usable output than less standardised text (Pierce 2018: 147).

1.6 Summing up

In this Chapter, the setting and background of the present study were described in detail, especially with regards to the South Tyrolean legal terminology, the public administration translation policies and the use of translation technologies and resources in the public administration's translation workflows. Several gaps have been identified (Sandrini 2019; De Camillis 2021) in the public administration's translation workflow, in particular with regards to translation technologies and resources. Possible development proposals advanced by Sandrini (Sandrini 2019: 391) in order to improve the translation workflows in South Tyrolean administrations include filling some of these gaps, in particular by: a) collecting an official translation memory based on already published texts in the LexBrowser collection and make it available as a TMX file on the LexBrowser website; b) adapting an MT system with local translation memories and terminology tailoring it to the local needs, and making it freely available online. The present work aims at partially filling these gaps: all Italian-German legal-administrative texts published in the LexBrowser will be collected and aligned, and a first attempt to adapt an MT system to the South Tyrolean legal-administrative texts will be made. The adapted MT system will then be evaluated, with particular focus on legal term translation, based on an ad-hoc terminology evaluation framework.

CHAPTER 2

2. MACHINE TRANSLATION

2.1 Introduction

Machine translation (MT) refers to “computerised systems responsible for the production of translations from one natural language into another, with or without human assistance” (Hutchins and Somers 1992: 3). In the last few years, MT quality has made huge progress, especially with the shift towards neural machine translation (NMT) approaches (see Section 2.3.2.3). Although some researchers claimed they achieved human parity in certain language pairs and domains (Hassan et al. 2018), this goal still seems far from being reached (Läubli *et al.* 2018). While MT cannot compete with professional translators with respect to translation quality, it is a valuable tool that can be used by translators to increase productivity, especially with certain domains and genres (Koehn 2020:36). Thanks to the significant progress achieved in MT quality, machine translation is being progressively adopted not only by a growing number of translators, LSPs and translation departments of companies and international organisations, but also by governments and public administrations.¹

This Chapter will be devoted to machine translation, with particular focus on issues related to terminology, which is of central importance in the present work and in legal-administrative settings. In particular, after briefly outlining the historical development of MT, the main MT architectures will be reviewed, with particular emphasis on neural machine translation. Moreover, after covering the topics of MT adaptation, data filtering, and MT use in public settings, related work on machine translation of legal-administrative texts and on terminology evaluation will be reviewed.

¹ MT as an aid for translators is usually implemented within a machine translation post-editing (MTPE) workflow, i.e., translators edit and correct machine translation output, either to obtain a merely comprehensible text (“light post-editing”) or to yield a product comparable to a product obtained by human translation (“full post-editing”) (ISO 18587:2017). Other MT applications include speech-to-speech machine translation (e.g., Skype Translator, machine interpreting), multimodal machine translation (e.g., image caption translation, subtitle translation) and sign language translation (Koehn 2020: 23–28).

2.2 Historical overview

Although the origins of MT can be traced back to the seventeenth century, when the first ideas of universal languages and mechanical dictionaries arose, the earliest practical suggestions were made in the first half of the twentieth century (Hutchins 2010: 1). In the 1930s, the first patents of automatic machines to assist in the translation of languages were filed in France and Russia by G. Artsrouni and P. Trojanskij. Artsrouni designed a storage device on paper tape which could also function as a mechanical multilingual dictionary by finding the equivalent of any word in another language (Hutchins 1995: 432–433; Hutchins 2010: 1). Trojanskij proposed a three-stage translation process with a first stage of human lemmatisation and syntactic analysis, a second stage of translation of base forms into another language by means of a mechanical dictionary, and a third stage of “post-editing” by a human being (Zarechnak 1979: 7–8).

The idea of machine translation was brought to general notice in 1949, with the publication of Warren Weaver’s *memorandum*, where prospects and possible research outlooks in MT were presented (Hutchins and Somers 1992: 5–6). Weaver’s *memorandum* paved the way for a period of active research in MT. In 1951, Bar-Hillel was appointed as the first MT researcher at MIT, where he held the first MT conference in 1952. In 1954, the Georgetown University and IBM held the first MT demonstration, by translating 49 Russian sentences into English, using a vocabulary of 250 words and just 6 grammar rules (Hutchins 1995: 433).

The declared aim of initial research in the field of MT was to reach fully automatic, high-quality machine translation (FAHQMT) for an unrestricted range of texts, but this goal was soon exposed as unrealistic and impossible to reach (Quah 2006: 7, 61). In 1964, it became clear that MT could not overcome the “semantic barrier” of natural language (Yngve 1964: 279). In 1966, ALPAC (Automatic Language Processing Advisory Committee) published a report which stated that human translation was faster, more accurate, twice as cheap as MT and that “there is no immediate or predictable prospect of useful machine translation” (ALPAC 1966: 32). Instead of further investing in MT research, it suggested that focus be shifted to developing tools to aid translators and to supporting basic research in computational linguistics (Hutchins 1995: 433). As a consequence of the ALPAC report, research in the field of MT almost

completely stopped for almost a decade in the USA (*ibid.*), whereas research groups in other countries (Canada, Germany, France) carried out further research on MT (Somers 2011). In the 1970s and 1980s, the first commercial MT systems were developed, such as Météo, developed by the University of Montréal to translate weather forecasts, Systran and Logos, which focused only on a few language pairs (Koehn 2020: 35). In the 1980s, the resurgence of MT research, as well as the commercialization of MT systems, led to increased public awareness of the importance and necessity of translation tools (Hutchins and Somers 1992: 9), which also brought about the development of several tools to be integrated into the “translator’s workstation” (Hutchins 2010: 10–11).

Whereas MT systems developed until then were *rule-based*, in the late 1980s and 1990s there was a shift towards *data-driven* or *corpus-based* approaches (see Section 2.3) (Hutchins 1995: 440; Koehn 2009: 17). The emergence of statistical MT, in particular, was ground-breaking and the approach gained full momentum by the year 2000, also due to the increase in computing power, data storage, availability of digital texts and the size of the Internet (Koehn 2009: 17–18). Starting from the launch of BabelFish in 1997, many similar online MT systems were made freely available in the following decade, progressively covering a larger number of language pairs and making MT a mass-market product (Hutchins 2011).

In the last few years, with the application of neural networks in the MT field, a paradigm shift from statistical machine translation to neural machine translation (NMT) has taken place (Stahlberg 2020: 54). Although they still suffer from some major limitations (Castilho *et al.* 2017), NMT systems have shown to perform better than SMT systems in terms of fluency and accuracy and represent today’s state-of-the-art in MT research (Luong and Manning 2015; Bentivogli *et al.* 2016; Toral and Sánchez-Cartagena 2017). Today, all major commercial MT systems are powered by neural systems (Stahlberg 2020: 2) with many MT providers also allowing for adaptation by means of own in-domain parallel corpora and glossaries.

2.3 MT architectures

From early rudimentary systems based on hand-crafted rules to today’s neural systems based on deep learning, various approaches and system architectures have been adopted in machine translation. MT architectures can be broadly grouped into:

- *rule-based* approaches: MT architectures based on dictionaries and linguistic rules. These include *direct* and *indirect* (*transfer* and *interlingua*) approaches;
- *corpus-based* (or *data-driven*) approaches: MT architectures that rely on large parallel corpora. These include *example-based*, *statistical* and *neural* MT systems.

In the following Sections, the architectures of various MT approaches developed will be discussed in greater detail.

2.3.1 Rule-based approaches

The first MT systems developed from the 1950s to the late 1980s were *rule-based*, i.e., they were not based on existing translations, but on monolingual and bilingual dictionaries and linguistic rules of different kinds (rules for syntactic analysis, lexical rules, rules for lexical transfer, rules for syntactic generation, rules for morphology, etc.) (Hutchins 1995: 440; Quah 2006: 70–71). Since such sets of rules were hand-crafted, a huge amount of human input was needed (Okpor 2014: 60). Sub-approaches in rule-based machine translation (RBMT) are the *direct* and *indirect* (*interlingua* and *transfer*) approaches, which can be best represented using the Vauquois (1968) triangle (Fig. 2).

Direct translation was the first approach employed in early MT systems (like the Georgetown/IBM system (1954), see Section 2.2) and is considered the “first generation” of MT systems (Quah 2006: 60, 69). Direct systems are essentially dictionary-based systems that carry out word-for-word translation of each source-language word into its target-language equivalent, without any stage of linguistic analysis (Quah 2006: 69; Q. Liu and Zhang 2015: 110). Direct MT systems are designed specifically for one particular pair of languages in one direction, and cannot properly handle idiomatic expressions, ambiguities or translations between unrelated languages (Hutchins and Somers 1992: 4; Quah 2006: 69–70; Okpor 2014: 161).

Indirect approaches (which include *interlingua* and *transfer* translation) are considered the “second generation” of MT systems and include a stage of syntactic or semantic analysis based on linguistic rules, in order to create abstract or intermediate language representations (Hutchins 1995: 431; Quah 2006: 63, 71). In the *interlingua* approach, in particular, translation is carried out in two stages: the source text is first converted into an abstract language-independent semantic representation and then “translated” into the target language(s) by means of dictionaries and grammar rules (Hutchins and Somers 1992: 4; Hutchins 2010: 3; Quah 2006: 71). In the *transfer* approach, instead, separate representations for source-language and target-language texts are produced, with the system moving from source text to source-language representation (analysis) and target-language representation (transfer) before producing the target text (generation) (Hutchins and Somers 1992: 4–5; Quah 2006: 71, 73–76). Whereas in direct translation systems rules for analysis, transfer and generation were not clearly separated, indirect systems and corpus-based systems feature various degrees of modularity, allowing for an independent adaptation and modification of the system’s components, data and programs (Hutchins and Somers 1992: 5; Quah 2006: 68–69, 192).

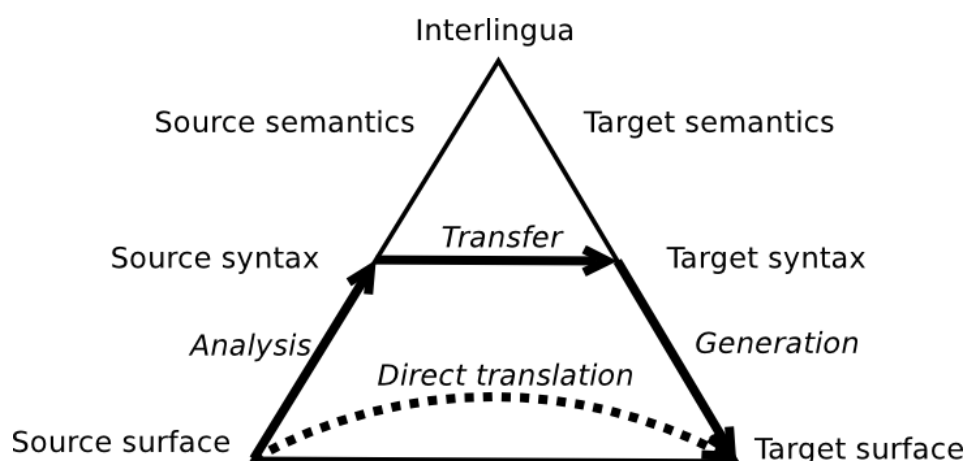


Figure 2: The Vauquois triangle (1968) representing the levels of analysis and generation in different rule-based MT approaches.

2.3.2 Corpus-based approaches

Since the late 1980s, new approaches based on parallel corpora (called *corpus-based* or *data-driven* approaches) have emerged and progressively replaced existing rule-based architectures (Koehn 2009: 17; Q. Liu and Zhang 2015: 108). Corpus-based MT

systems do not use any linguistic rules and are based on algorithms that analyse and extract translation examples from large amounts of raw data in the form of parallel corpora (Quah 2006: 64; Okpor 2014: 160–162). Corpus-based approaches include example-based machine translation (EBMT), statistical machine translation (SMT) and, more recently, neural machine translation (NMT) (Wang *et al.* 2021: 2).

2.3.2.1 Example-based machine translation

Example-based machine translation systems (EBMT, also called *memory-based* MT systems) were first proposed in 1984 and were based on the assumption that translation involves the recall of similar, already translated examples (Nagao 1984; Hutchins 1995: 440). The approach consists in selecting and extracting equivalent translation segments from an aligned parallel corpus. More precisely, after having retrieved the examples from the parallel corpus, the input sentence is decomposed into fragments in order to match example fragments. Each matched source fragment is then translated according to the word alignment between source and target examples. In the final target sentence recombination stage, the translation of the source fragments is assembled into the target sentence (Q. Liu and Zhang 2015: 112–113; Hutchins 1995: 440–441). Although it was a major difficulty to produce fluent and grammatical output by means of re-combination of target language examples in the form of short phrases, the main advantage of the EBMT approach, compared to RBMT, was that its output displayed a good level of idiomaticity, since text fragments were extracted by actual human translations (Hutchins 2010: 12).

2.3.2.2 Statistical machine translation

Statistical machine translation (SMT) is “a machine translation paradigm that generates translations based on a probabilistic model of the translation process, the parameters of which are estimated from parallel text” (Y. Liu and Zhang 2015: 201). The first statistical MT system was developed by IBM in 1989 and SMT remained the state-of-the-art and most studied approach until the advent of neural MT systems (Way 2020: 311).

The idea at the basis of SMT is to generate a translation model, i.e., to mathematically model the probability of a target sentence being the translation of a given source sentence. The translation process then consists in searching an optimal target sentence with the highest translation probability from the space of all possible target sentences for a given sentence by means of a decoding algorithm. The final translation output is finally “refined” by means of a target language model. The basic components of a statistical machine translation architecture are therefore a translation model, a language model and a decoder. Whereas the translation model is trained on parallel corpora and ensures that the MT system produces a target sentence corresponding to the source sentence, the language model is trained on target-language monolingual corpora and ensures a fluent and grammatically correct output from the MT system (Q. Liu and Zhang 2015: 113; Okpor 2014: 163; Garg and Agarwal 2019: 2–3).

Translation models and, accordingly, SMT architectures can be classified, according to the language units used, into *word-based*, *phrase-based* and *syntax-based* MT systems (Q. Liu and Zhang 2015: 113). Word-based SMT models, implemented by early SMT systems, calculate sentence translation probability based on word-to-word translation tables (Brown *et al.* 1993; Y. Liu and Zhang 2015: 202–204). Phrase-based (PBMT) models, instead, process phrases (usually 3-grams) instead of single words, and are therefore based on phrase tables with phrase-to-phrase translation probabilities. The advantage of PBMT systems is the ability to take local context into consideration and to handle word insertion and deletion and the translation of idioms, therefore outperforming word-based models. However, phrase-based models fail to effectively process long-distance dependencies and to achieve effective global reordering (Koehn *et al.* 2003; Q. Liu and Zhang 2015: 114; Y. Liu and Zhang 2015: 205–207). Syntax-based systems, finally, take advantage of dependency parsing trees on the source and/or target side. The main disadvantages of syntax-based models are the scarce availability and accuracy of parsers and the bigger size of the model compared to phrase-based systems, with consequently higher memory requirements and significantly lower translation speed (Williams *et al.* 2016; Q. Liu and Zhang 2015: 114; Y. Liu and Zhang 2015: 207–210).

Although being the state of the art in MT for almost two decades, statistical machine translation had some important drawbacks. Its inability of modelling and handling long-distance dependencies between words, for example, made the translation quality of SMT far from satisfactory in terms of fluency (Tan *et al.* 2020: 5).

2.3.2.3 Neural machine translation

The first MT models based on artificial neural networks were proposed as early as 1997, but for several years the implementation of neural networks in MT was scattered and did not achieve state-of-the-art results due to missing data and, especially, to low computational resources available at the time (Koehn 2020: 39; Stahlberg 2020: 1). In 2014, the first end-to-end² translation models based entirely on neural networks (Bahdanau *et al.* 2014; Sutskever *et al.* 2014) were proposed and were referred to using the term “neural machine translation” (Wang *et al.* 2021: 2). In just a few years, NMT outperformed SMT, achieving state-of-the-art performance on various language pairs (Junczys-Dowmunt *et al.* 2016; Bentivogli *et al.* 2016; Wu *et al.* 2016) and becoming the *de facto* paradigm in MT (Stahlberg 2020: 54; Tan *et al.* 2020: 5; Zhang and Zong 2020: 2).

NMT is based on artificial neural networks, i.e., advanced deep learning algorithms which loosely mimic the functioning of a human brain. Neural networks are organized in layers (usually an input layer, a number of hidden layers, and an output layer, see Fig. 3) of basic units called neurons or nodes, which are designed to behave similarly to a neuron in the brain. Each node of the neural network combines inputs, an activation function and an output value, and is connected to all nodes of the previous and following layer with weighted connections. If the computed value within a node exceeds a certain threshold according to the activation function, the node “fires” (i.e., it is activated) and passes the value on to other nodes in the network (Müller *et al.* 1995: 13–17; Gurney 1997: 12–16; Aggarwal 2018: 1–20; Koehn 2020: 31, 67–79).

² NMT models use a single large neural network to model the entire translation process, as opposed to late RBMT and SMT systems, which are composed by several separate components.

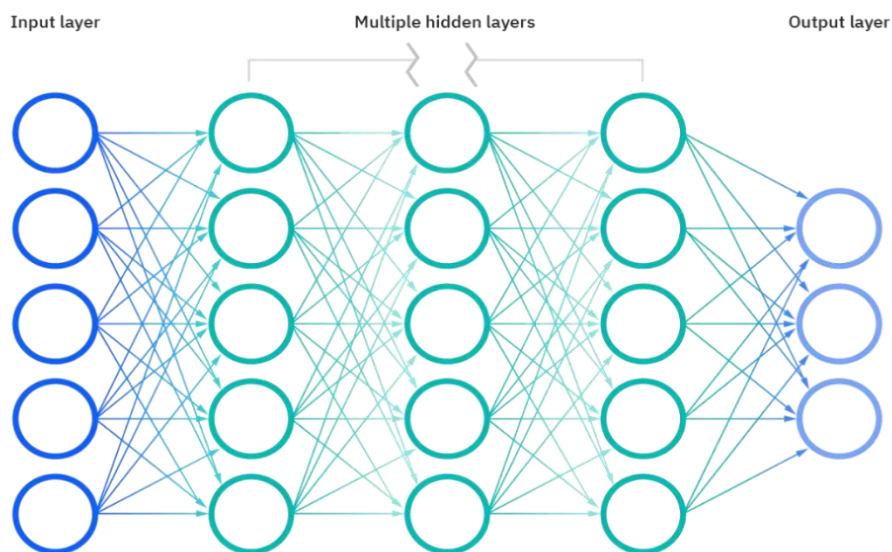


Figure 3: The basic structure of a deep neural network.

Neural machine translation systems are sequence-to-sequence models based on an end-to-end *encoder-decoder* architecture (Sutskever *et al.* 2014; Cho *et al.* 2014). The encoder is a neural network which takes a source-language sentence and maps each unit³ into a low-dimensional real-valued vector (a word embedding (Mikolov *et al.* 2013)) and encodes the sequence of vectors into distributed semantic representations. The encoder neural network, in turn, decodes these numerical representations and generates the target-language sentence one output word at a time (Zhang and Zong 2020: 2; Wang and Sennrich 2020: 2). Neural networks used to build encoders and decoders within NMT systems can be roughly divided into three categories (Tan *et al.* 2020: 8): recurrent neural networks (RNN), convolutional neural networks (CNN) and self-attention networks (SAN).⁴

Early NMT architectures (Kalchbrenner and Blunsom 2013; Cho *et al.* 2014; Sutskever *et al.* 2014) adopted a fixed-length approach, i.e., the size of source-sentence representations was fixed regardless of the length of the sentence, and implemented

³ Sentences are split into tokens, characters or subword units (usually based on byte-pair encoding (BPE) (Sennrich *et al.* 2016b)). Subwords are currently the most commonly implemented translation unit in NMT (Stahlberg 2020: 28).

⁴ For a more in-depth review and comparison of different NN architectures implemented in NMT see, f.e., (Stahlberg 2020; Tan *et al.* 2020; Wang *et al.* 2021; Garg and Agarwal 2019; Zhang and Zong 2020; Koehn 2020).

RNN architectures. One drawback of such approach, however, is performance degradation when translating longer sentences (Tan *et al.* 2020: 7). Later architectures, therefore, adopted a variable-length approach, in particular by implementing an attention mechanism (Bahdanau *et al.* 2014). The introduction of the attention mechanism is considered a milestone in NMT architecture research (Tan *et al.* 2020: 8), as it improves performance by allowing the decoder to selectively focusing on sub-parts of the source sentence during translation (Bahdanau *et al.* 2014; Luong and Manning 2015). In the following best-performing models, RNNs were progressively replaced by CNNs (Tan *et al.* 2020: 8). The current state-of-the-art NMT architecture in terms of quality and efficiency is the Transformer model (Vaswani *et al.* 2017). By relying entirely on self-attention networks (SAN), the Transformer generates context-sensitive word representations which depend on the whole source sentence and is able to “draw global dependencies between input and output” (Vaswani *et al.* 2017: 2), therefore improving overall fluency and quality (Stahlberg 2020: 15).

2.4 Parallel corpora for MT training

As a *data-driven* approach, the training of statistical and neural machine translation systems relies on parallel corpora and, therefore, often suffers from data sparsity⁵ problems (Koehn 2020: 39). Neural machine translation, in particular, has been found to be particularly data hungry, i.e., it requires a significantly larger quantity of parallel sentences for training in order to achieve satisfactory results compared to SMT approaches (Zoph *et al.* 2016: 1; Koehn and Knowles 2017: 28). For that reason, a number of approaches have been proposed in order to increase the size of available training data (data augmentation), such as synthetic corpus generation (usually back-

⁵ This phenomenon is linked to the Zipf’s law, one of the few mathematical laws in natural language. According to the Zipf’s law, given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Non-lexical words are the most frequent, whereas lexical words are at the tail of the distribution. Huge corpora are needed to mitigate this distribution.

translation, i.e., machine-translating monolingual target corpora into source language to increase training data size) (Sennrich *et al.* 2016a).⁶

Moreover, what is central to achieve good MT performance is not only training data *quantity*, but also *quality*. This holds true particularly for neural machine translation models, which have been proved to be by far more sensitive to noise in training data than statistical systems (Chen *et al.* 2016; Koehn and Knowles 2017). Therefore, the data preparation, cleaning and filtering stage is crucial in MT training: it makes it possible to reduce the amount of redundant data in the training set, to avoid the “garbage in, garbage out” problem and, accordingly, to improve the quality of the MT system.

Considerable amount of research has been carried out on filtering out noise in parallel data, including both rule-based and supervised/unsupervised machine learning and deep learning approaches (Koehn *et al.* 2020). Since corpus filtering is a significant stage carried out in the present dissertation to achieve MT adaptation (see Section 3.3), in the following subsection, noise categories in parallel corpora and the rule-based filtering approaches commonly adopted in parallel corpus cleaning for (N)MT training will be briefly reviewed.

2.4.1 Cleaning noise in parallel data

In the present Section, the main categories of noise occurring in parallel corpora and their influence⁷ on NMT systems are listed. Moreover, the main approaches based on deterministic rules commonly adopted to discard noisy sentence pairs when cleaning parallel datasets for MT training are briefly described.⁸ Categories of noise in parallel corpora include:

⁶ In order to bypass the data sparsity problem and the reliance on parallel corpora for NMT training, a fair amount of recent research has also been focusing on unsupervised neural MT (UNMT), typically based on language model pre-training and subsequent fine-tuning using back-translated sentences (Marie and Fujita 2018; Lample *et al.* 2018; Conneau and Lample 2019).

⁷ As observed by Khayrallah and Koehn (2018).

⁸ The review of the most common pre-filtering operations in MT dataset cleaning is based on the contributions of the participants of the WMT18, WMT19 and WMT20 Shared Tasks on Parallel Corpus Filtering (Koehn *et al.* 2018; Koehn *et al.* 2019; Koehn *et al.* 2020). Participants also adopted more

- **MISALIGNED SENTENCES:** Sentence pairs in which source and target segments do not match, due to faulty document or sentence alignment or segmentation issues. Expectedly, such noise has a negative influence on MT quality of up to -1.9 BLEU (Khayrallah and Koehn 2018: 78). Filtering operations commonly adopted to identify and discard misaligned sentence pairs include:
 - **Digit matching:** Matching digits between source and target segments and discarding sentence pairs where these digits differ at all.
 - **Length-based filtering:** Removing sentence pairs that have widely varying lengths in terms of tokens or characters. This is achieved either by: a) discarding sentence pairs with a difference in terms of tokens (Kurfalı and Östling 2019) or characters (Hangya and Fraser 2018) exceeding a given threshold; b) computing the ratio of source/target tokens (Lu *et al.* 2018) or characters (Pinnis 2018) and discarding sentence pairs exceeding a given threshold; c) taking into account average length ratio and variance of all sentences in the corpus and discarding outliers (Jalili Sabet *et al.* 2016) or pairs exceeding a certain number of standard deviations from the overall average length ratio (Gupta *et al.* 2019; P. Chen *et al.* 2020)
 - **Hunalign score:** Filtering out segments based on their Hunalign alignment score.⁹
- **MISORDERED WORDS:** Disfluent language may be due to poor human translation, not post-edited machine translation, or heavily specialized language. According to the experiments carried out by Khayrallah and Koehn (2018: 78), this kind of noise has negative influence on MT quality (up to -1.7 BLEU).

sophisticated approaches based on machine learning techniques, e.g., assessing sentence filtering using scoring functions (Koehn *et al.* 2018; Koehn *et al.* 2019) or even framing the problem as a classification task (Koehn *et al.* 2020). Since the filtering operations carried out in the present work will be mainly rule-based, machine learning approaches to corpus cleaning and filtering have been left out of the scope of the present review.

⁹ Hunalign (Varga *et al.* 2005) is an algorithm that automatically aligns bilingual texts at the sentence level. The hunalign alignment score is the similarity score assigned by the algorithm to each aligned pair of sentences.

However, it must be noted that the study was carried out on a case of extreme disfluency, namely a dataset with sentences reordered randomly, which is hardly a realistic scenario. Such type of noise can hardly be detected by means of rule-based approaches.

- **WRONG LANGUAGE:** Sentence pairs may be in languages other than the ones expected, causing a quality loss of up to -2.2 BLEU (Khayrallah and Koehn 2018: 78). Several algorithms for language detection are freely available online and can be used to detect source and target language and filter out sentence pairs not matching the required languages.
- **UNTRANSLATED SENTENCES:** Sentences with identical source and target. Khayrallah and Koehn found that this particular type of noise has a “catastrophic impact on neural machine translation, leading it to learn a copying behavior that it then exceedingly applies” (2018: 74), causing a huge MT quality loss of up to 24.0 BLEU.
- **HIGHLY SIMILAR SOURCE-TARGET:** Just like untranslated sentences, sentence pairs with highly similar source and target are a typology of noise which may have a negative influence on MT quality. Approaches to filter out this kind of noise include: a) computing BLEU similarity score (Song *et al.* 2014); b) computing Levenshtein’s edit distance and edit distance ratio between source and target (Lu *et al.* 2018); c) checking that more than half of the source words are not present in the target segment (Pinnis 2018).
- **TOO LONG AND TOO SHORT SENTENCES:** Sentences which are very long or very short are usually discarded according to various parameters, such as minimum and maximum token/character thresholds or the number of standard deviations from the average segment length (Gupta *et al.* 2019). Khayrallah and Koehn (2018: 78) actually found that very short sentences (≤ 2 tokens) have only a small negative influence on MT quality (up to just -0.7 BLEU), whereas short sentences (3-5 tokens) could even lead to a small increase in MT quality (up to +0.8 BLEU). Minimum and maximum token/character thresholds vary significantly among different filtering methods.
- **WRONG CHARACTERS:** In a noisy corpus, segments containing unwanted characters need to be filtered out. Ash *et al.* (2018) defined a list of characters

considered “acceptable” and only keep sentence pairs containing those characters. Pinnis (2018) removes sentence pairs when words contain question marks between letters, indicating encoding corruption in data.

- **NON-ALPHABETICAL SENTENCES:** Segments which are probably not very useful (or even harmful) for the NMT system such as, for instance, segments composed only or mostly by punctuation, digits or whitespaces (Gupta *et al.* 2019). Such segments are usually filtered out if a) a sentence is made up of just punctuation, digits and whitespaces; b) the ratio between digits/punctuation and actual alphabetical characters exceeds a certain threshold (Rikters 2018); c) one side in the sentence pair contains significantly more non-alphabetical characters than the other side (Rikters 2018).
- **DUPLICATE SENTENCE PAIRS:** Identical sentence pairs can be deduplicated. Pinnis (2018) applies more sophisticated deduplication filtering by removing almost-duplicates based on normalized sentence pairs (where whitespaces and punctuation are removed, digits are replaced with a placeholder, and everything is lowercased). This makes it possible to discard more redundant data than a simple deduplication of just identical source-target pairs.
- **HALF-EMPTY SENTENCE PAIRS:** Sentence pairs where the source or target sentence is empty or contains just whitespaces. This type of noise can be due to misalignments or 1:0 alignments, or a consequence of other cleaning/normalization operations carried out on the corpus. Such sentence pairs can be easily identified and filtered out.
- **LONG STRINGS:** Sentence pairs containing any token longer than a given character threshold (for instance, 30 characters (Kurfalı and Östling 2019) or 50 characters (Pinnis 2018)), indicating a possible missing whitespace that resulted in merged words.
- **SHORT OR LONG WORDS:** Sentence pairs where either sentence has an average word length lower or higher than a given threshold.
- **INCONSISTENCIES IN SOURCE AND INCONSISTENCIES IN TARGET:** Sentence pairs where the same source and different translations and, conversely, multiple source sentences aligned to the same target translation.

- **INCONSISTENCY OF SPECIAL TOKENS:** Sentence pairs where there are inconsistencies of URLs, digits, e-mail addresses between source and target.

2.5 Machine translation for institutional translation

Machine translation is used not only by freelance translators, LSPs and multilingual companies, but it is also widely adopted in institutional and government settings (Dillinger and Lommel 2004: 4; Nurminen and Koponen 2020: 152, 156–159). This is the case especially within large international organisations (EU, UN, etc.) as well as governments and administrations of bilingual or multilingual countries (Switzerland, Canada, etc.), which have large needs for multilingual legislation and documents, but also public sectors of non-multilingual countries, which have adopted MT to improve information accessibility (Nurminen and Koponen 2020).

The European Union was a pioneer in this context, starting to develop MT systems as early as 1976 (Hutchins 1995: 437). In 2013, MT@EC, a statistical MT system translating among all official EU languages was released and made available for staff of the EU institutions and bodies, public administrations in the EU member states and members of the European Masters in Translation network (Mai 2016). In 2017, the EU released eTranslation, a neural machine translation system built upon MT@EC. Today, eTranslation translates between all 24 official EU languages plus Icelandic, Norwegian, Russian and simplified Chinese, and is available for staff of the EU institutions and bodies as well as for European SMEs.¹⁰

Other multinational organisations such as PAHO (Vasconcellos and Leon 1985) and the United Nations have also implemented SMT and NMT systems within the translation workflows of their organisation and specialised agencies (WHO, WIPO, IMO, ITU, ILO, WTO, TGF) (Pouliquen *et al.* 2011; Pouliquen *et al.* 2012; Pouliquen *et al.* 2013; Junczys-Dowmunt *et al.* 2015; Pouliquen 2017).

In officially multilingual countries (such as Canada and Switzerland), where a large number of documents is drafted or translated into more than one language on a

¹⁰ eTranslation, (<https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation>).

daily basis, MT has been implemented in the institutional translation workflow at different levels and extents. In Canada, which has a long history in developing MT systems for public uses (see the Météo system, Section 2.2), MT is widely used within the Canadian Translation Bureau (Seguin 2021). In Switzerland, in 2019 DeepL Pro was adopted as a support for employees and translators of the federal administration (Schweizerische Eidgenossenschaft 2019).

In non-multilingual countries, too, MT has been deemed a useful potential tool to improve information accessibility and customized MT solutions have been developed and/or implemented for use at institutional level. For instance, customized MT systems have recently been developed by Tilde for the governments of Latvia and Lithuania in the Latvian/Russian, Latvian/English, Lithuanian/French and Lithuanian/English language pairs (Vasiljevs *et al.* 2014; Skadins *et al.* 2020). In the framework of the PRINCIPLE project, MT systems for public sector users have been developed in Croatia, Iceland, Ireland, and Norway (Way *et al.* 2020). Projects for the development and implementation of MT systems in public administrations or agencies have also been carried out in Japan (Miyata *et al.* 2016) and Sweden (Sågvall Hein and Ekholm 2020).

In multilingual regions where minority languages officially co-exist with the majority language (for instance, the Basque Country and Catalonia),¹¹ MT systems have been created to translate between the majority language and the minority languages. In particular, in Catalonia, a significant number of MT systems (both commercial and open-source, like the Apertium MT platform (Armentano Oller *et al.* 2007)) have been created to translate from and into Catalan and Aranese, the region's minority languages. Moreover, attempts to integrate MT into the translation workflow of the Catalan public administration have been underway for around 20 years (De Camillis 2021: 120–121). In the Basque Country, several tools have been developed for machine translation between Castilian and Basque, some of which are made available online to the general public. Although a number of experiments on the deployment of MT have been carried

¹¹ As analysed by De Camillis (2021), the Basque Country and Catalonia have linguistic and political situations comparable to South Tyrol and, therefore, can serve as a means of comparison to the South Tyrolean region with regards to their language policies and several aspects thereof, including machine translation.

out in the last years, the Basque public administrations have not yet systematically integrated MT in their translation workflow (De Camillis 2021: 147).

A final example worth mentioning, since it represents an attempt to propose an MT system suited to the legal terminology specific to a German-speaking legal system (which coincides with the aims of the present work), is the creation of an NMT system tailored to Austrian German for the EU Council Presidency Translator for Austria, based on Austrian German Language Resources (Heinisch and Lušický 2020).

2.6 NMT and terminology

Despite the significant improvements achieved by NMT, which outperformed statistical approaches in the last few years, neural systems suffer from performance degradation when exposed to new domains¹² or applied in multi-domain scenarios, “to the point that they completely sacrifice adequacy for the sake of fluency” (Koehn and Knowles 2017: 28). Quality deterioration for specific domains is due to the distance between the target domain and the domains which the MT systems are trained on and mainly causes issues in handling domain-specific terminology (Farajian *et al.* 2018; Dinu *et al.* 2019). However, since the correct translation of domain-specific terminology is of central importance in translation, several approaches have been proposed to tackle the issue of domain and terminology adaptation.

Whereas in statistical machine translation it was relatively easy to integrate domain-specific terminology and to force the translation of certain words and phrases due to the nature of the underlying MT architecture,¹³ introducing hard translation

¹² A customary definition of “domain” in MT is given by Koehn (2020: 239): “[A] collection of text with similar topic, style, level of formality, etc. In practical terms, however, it typically means a corpus that comes from a specific source”. This general definition allows for different levels of granularity in defining a domain: research on MT domain adaptation mainly deals with broader domains (e.g., medical, legal, IT, literature, news, subtitles, etc.) but also with narrowed down sub-domains (Kocmi 2019:6). In the present work, we will broadly deal with the legal domain in the South Tyrolean German variety. Assuming a “source-based” definition of domain like the one cited above, we can delimit the domain more specifically to all South Tyrolean legal-administrative texts (laws, decrees, regulations, etc.) published in the LexBrowser database. The domain considered has not been restricted to a particular legal sub-domain (e.g., environmental law, road law, procedural law, etc.) due to an implicit lack of data pertaining to each single sub-domain.

¹³ For an in-depth review of terminology integration approaches in SMT see Pinnis (2015).

constraints in NMT is not a trivial task (Chu and Wang 2018; G. Chen *et al.* 2020). One of the most prominent approaches to inject terminology in NMT systems is constrained decoding, whereby terminology is inserted as a set of constraints at decoding time (Chatterjee *et al.* 2017; Hokamp and Liu 2017; Post and Vilar 2018). More recent approaches, instead, are based on in-line term annotations (Dinu *et al.* 2019; Exel *et al.* 2020), or placeholders complemented with part-of-speech and morphological information both in source and target sides (Michon *et al.* 2020).

Adaptive approaches to tailor a NMT system to a specific domain or multiple domains, instead, propose to tune an existing MT system, trained on generic data or parallel data pertaining to other domains, to the required domain (Farajian *et al.* 2018: 150). The conventional domain adaptation approach is called *fine-tuning* (Luong and Manning 2015) and consists in training an MT system on a large generic out-of-domain parallel corpus, then tune its parameters on a smaller in-domain corpus. However, fine-tuning on in-domain data can lead to a risk of overfitting¹⁴ due to the small size of the in-domain corpus, causing a “catastrophic forgetting” of the general domain knowledge (Koehn 2020: 253). To address this problems, other domain adaptation approaches have been proposed, including *mixed fine-tuning* (fine-tuning on a mix of in-domain and out-of-domain data, (Chu *et al.* 2017)) and regularization techniques like *tuneout* (Miceli Barone *et al.* 2017), among others.¹⁵

The domain adaptation methods described above are applied at different moments before testing. Farajian *et al.* (2017), instead, proposed a technique called *instance-based adaptation*,¹⁶ which makes it possible to achieve on-the-fly multi-domain adaptation at translation time, based on a set of sentences in the pool of parallel data that are similar to the sentence to be translated.¹⁷ Instance-based adaptation has showed significant improvements over generic systems in multi-domain settings, also with

¹⁴ Overfitting happens when a model is overly adapted on the in-domain data, to the point that it yields poor performance on any other data.

¹⁵ For more detailed reviews of domain adaptation methods in NMT see Chu and Wang (2018), Koehn (2020:239-261) and Saunders (2021).

¹⁶ Based on Hildebrandt *et al.* (2005) and Li *et al.* (2016).

¹⁷ Further details about instance-based adaptation, which is the adaptive approach behind ModernMT, are given in Section 3.4.2.

regards to terminology translation (Farajian *et al.* 2017; Farajian *et al.* 2018).

Today, adaptation by means of parallel corpora and/or bilingual terminology to create user-specific NMT systems on top of existing state-of-the-art baseline models is offered by a number of commercial MT providers, including, but not limited to, Google's AutoML Translation,¹⁸ Microsoft Custom Translator,¹⁹ ModernMT,²⁰ DeepL,²¹ Amazon Translate,²² KantanMT,²³ Systran.^{24,25}

2.7 Evaluation of machine translation quality

The need to assess the quality of MT systems by analysing their output has existed since the early developments of MT. Since then, evaluation of MT systems has been studied extensively and is considered a crucial, yet challenging, task in MT research (Castilho *et al.* 2018: 24; Koehn 2020: 41; Han *et al.* 2021a).

MT quality can be evaluated either manually or by means of automatic evaluation metrics, which compare the output of an MT system (translation *hypothesis*) with one or several *reference* human translations (Koehn 2009: 217–232). Both of these approaches exhibit advantages and disadvantages, also according to the aims of the MT evaluation procedure (Castilho *et al.* 2018: 25–26). Although manual MT evaluation is more useful to evaluate complex linguistic phenomena and focus on certain error types with different levels of granularity, it is a slow, complex, expensive and subjective procedure. Moreover, it is not tunable and not reproducible, and inter-annotator

¹⁸ <https://cloud.google.com/translate/auttml/docs> (last accessed: 25/08/2021).

¹⁹ <https://www.microsoft.com/it-it/translator/> (last accessed: 25/08/2021).

²⁰ <https://www.modernmt.com/> (last accessed: 25/08/2021). ModernMT does not allow direct terminology integration.

²¹ <https://www.deepl.com/it/translator> (last accessed: 25/08/2021). DeepL does not allow corpus-based domain adaptation. The glossary function is limited to the EN<>DE, EN<>FR and EN<>SP language pairs.

²² <https://aws.amazon.com/it/translate/> (last accessed: 25/08/2021).

²³ <https://www.kantanai.io/> (last accessed: 25/08/2021).

²⁴ <https://www.systransoft.com/> (last accessed: 25/08/2021).

²⁵ For an overall view of today's available commercial MT systems and their respective functionalities, see Custom.MT's MT Comparison Tool. Available at <https://custom.mt/mt-tech/> (last accessed: 25/08/2021).

agreement can be an additional issue (Popović 2018; Castilho *et al.* 2018; Han *et al.* 2021a). A valuable alternative is therefore represented by automatic evaluation, which is faster, cheaper, more objective, consistent and requires minimal human intervention (Popović 2018: 130). Automatic quality assessment, however, does not readily indicate the type or severity of the problems of the MT output, it is less comprehensive, less granular and has limited ability to assess syntactic or semantic equivalence. Moreover, comparing the MT output with human reference translations is intrinsically a biased operation with a subtle element of subjectivity and variability, since each machine-translated sentence is compared only to one among all possible correct translations of a given source sentence (Castilho *et al.* 2018: 25–26; Kocmi *et al.* 2021; Han *et al.* 2021a).

Several frameworks for manual MT evaluation have been developed over time and mainly aim at evaluating MT output *adequacy*²⁶ and *fluency*²⁷ at the sentence level (Koehn 2009: 217–220; Castilho *et al.* 2018: 17–18). Human evaluation taxonomies can have different levels of granularity and scoring methods, but common error categories include grammar, syntax, lexicon, omission, addition, style and terminology.²⁸ The main taxonomies developed for both human translation and machine translation over time include, but are not limited to, the LISA QA,²⁹ SAE J2450,³⁰ and, more recently, the DQF (O’Brien *et al.* 2011) and MQM (Lommel *et al.* 2014a) frameworks, which were later harmonised into DQF-MQM (Lommel 2018).

Regarding automatic MT evaluation, a very large number of metrics have been developed in MT research – according to Marie *et al.* (2021), as many as 108 new metrics have been proposed in the last decade. However, the most reported metric in papers involving MT experiments is by far BLEU, a *de facto* standard in MT research (*ibid.*). BLEU (Papineni *et al.* 2002) computes precision scores based on n-gram (sized

²⁶ *Adequacy* (or accuracy) is defined as “the extent to which the translation transfers the meaning of the source-language unit into the target” (Castilho *et al.* 2018: 18).

²⁷ *Fluency* 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)” (*ibid.*).

²⁸ For a review of error typologies in manual MT quality evaluation see Popović (2018).

²⁹ No public reference currently available (Lommel *et al.* 2014b: 457)

³⁰ <https://www.sae.org/standardsdev/j2450p1.htm> (last accessed: 26/08/2021)

from 1 to 4) overlaps and a sentence brevity penalty factor. Limitations of BLEU are well-known: BLEU disregards recall, its scores are difficult to interpret, it penalizes differences in word order and inflection from the reference sentence, and it doesn't correlate highly with human judgements (Koehn 2009; Way 2018; Mathur *et al.* 2020; Kocmi *et al.* 2021). As Post (2018) demonstrated, BLEU is not a single metric and doesn't always allow comparability, since its scores depend on its internal parameters and on the pre-processing steps (e.g., tokenization, normalization, etc.) carried out on the hypothesis and reference sentences. Therefore, he proposed a standard tool (SacreBLEU)³¹ to compute BLEU scores and achieve correct comparability across studies. However, SacreBLEU is still not widely adopted in the MT community (Marie *et al.* 2021).

NIST (Doddington 2002) is a close derivate of BLEU, inasmuch as it focuses only on n-gram precision and disregards recall. However, differently from BLEU, which assigns equal weight to all n-grams, NIST takes n-gram frequency into account when assigning weights. METEOR (Banerjee and Lavie 2005) is based on unigram matching, precision, recall and takes into account morphological variants and synonyms as well. LEPOR (Han *et al.* 2012) combines precision, recall, enhanced sentence-length penalty and n-gram based word order penalty. The hLEPOR (Han *et al.* 2013) enhanced variant achieved high correlations with human judgements and has been widely adopted in MT and NLP (Han *et al.* 2021b).

Metrics based on edit distance (rather than n-grams) include Word Error Rate (WER) and Translation Error Rate (TER). WER (Nießen *et al.* 2000) computes the minimum number of edits at the word level (insertions, deletions, substitutions) needed to change the translation hypothesis in order to match the reference exactly. TER (Snover *et al.* 2006) is a derivate of WER, but additionally takes into account shifts of word sequences. Recently, Alam *et al.* (2021) proposed TER_m, a terminology-biased modified version of TER, which penalizes errors that concern the terminology tokens more than other tokens.

³¹ <https://github.com/mjpost/sacrebleu> (last accessed 26/08/2011)

With state-of-the-art NMT relying mainly on subword segmentation approaches, character-based evaluation metrics have progressively been deemed “more appropriate” (Way 2018: 171) and have shown to achieve higher correlations with human evaluations than word-based metrics (Lardilleux and Lepage 2017: 146). The main character-based metrics include chrF (Popović 2015), based on character n-gram F-score,³² CharacTER (Wang *et al.* 2016) and CHARCUT (Lardilleux and Lepage 2017).

Recently, new evaluation metrics based on deep learning approaches have been proposed. By attempting to measure semantic similarity between source and target sentences, such models overcome some limits of traditional string-based approaches. BERTscore (Zhang *et al.* 2020) computes context word embeddings and pairwise cosine similarity between representations of a hypothesis and a reference translation. COMET (Rei *et al.* 2020) is based on cross-lingual language models and has achieved the highest correlation with human judgements in the latest benchmarks (Kocmi *et al.* 2021).

Recent recommendations suggest a shift from the exclusive use of BLEU score (Kocmi *et al.* 2021) or, at least, to compute it using the SacreBLEU tool to allow comparability (Marie *et al.* 2021). In particular, the use of pretrained models (like COMET) are recommended, as they achieve the highest correlations with human judgements. For language pairs where no pretrained model is available, character-based metrics (like chrF, the best performing string-based method) are suggested (Kocmi *et al.* 2021). Moreover, the use of significance tests³³ to corroborate quality improvements should become common practice in the MT community (Marie *et al.* 2021). Finally, Marie *et al.* (2021) introduced a scoring method for papers that rely on automatic metric scores for evaluating translation quality. The aim of the scoring method is to assess the trustworthiness of an automatic evaluation performed in an MT paper by assigning a

³² In particular, the chrF3 variant, with recall having 3 times more weight than precision, has achieved promising results (Popović 2015: 393).

³³ Statistical significance testing is a “standard methodology designed to ensure that experimental results are not coincidental” (Marie *et al.* 2021). In MT, it is carried out to assess whether a difference in metric scores between MT systems is statistically significant, i.e., not due to random chance. The prevalent methods in MT are the *paired bootstrap resampling test* (Koehn 2004) and the *approximate randomization test* (Riezler and Maxwell 2005).

score from 0 to 4 according to best practices related to evaluation metrics, significance testing and comparison with previous work.

2.8 Machine translation of legal-administrative texts

In the literature, a number of studies have focused on the evaluation of machine translation of texts in the legal domain in several language pairs, including the language pair of interest in this work (Italian-South Tyrolean German). In many of these studies, the focus was particularly on how MT systems could handle legal terminology, although a specific legal terminology evaluation framework has not yet been developed for this language pair. It must be noted that many of these studies were carried out before NMT was introduced in 2015.

Among studies that do not deal with the Italian-German language pair, Yates (2006) evaluated the accuracy of Babel Fish in translating legal information by comparing MT translations of law-oriented texts in Spanish and German to professional translations, concluding that Babel Fish is not appropriate for most uses in law libraries. Kit & Wong (2008) carried out an automatic evaluation of EU and UN legal texts translated from various languages into English in order to compare six online MT systems. Farzindar & Lapalme (2009) investigated the MT quality of Canadian court judgements in the English-French language pair by carrying out an evaluation based on edit distance and post-editing operations. Killman (2014) manually evaluated the accuracy of the terminological and phraseological translation choices provided in English by Google Translate for a selection of terminology items extracted from Spanish judgement summaries, finding that GT could “translate accurately vocabulary taken from a voluminous legal text aimed at expert readers in a little over 64% of the cases” (2014: 96). Mileto (2019) carried out a case study to evaluate to what extent translators may improve the translation quality of legal texts when working with a NMT system integrated into a CAT tool, carrying out a manual error analysis of the outputs yielded by MT@EC, SDL Machine Translation Cloud and Google Translate.

Among studies that deal specifically with legal NMT in the Italian-German language pair, Heiss and Soffritti (2018) analysed the output of NMT on an excerpt of a Provincial Law issued in South Tyrol, machine-translated from Italian into German using DeepL. They found that the NMT system yielded an understandable translation,

which was correct with regards to morphology and syntax, whereas the main errors identified regarded the legal terminology specific to South Tyrol. Wiesmann (2019) evaluated the Italian to German MT quality of various legal texts, including an excerpt from a South Tyrolean provincial law, a legal essay, a power of attorney, a notarial real estate sale contract, a statement of claim and a civil court judgement. The systems used to generate MT output were DeepL and a combination of NMT systems (Google Translate, DeepL Translator and Microsoft Translator) integrated in the MateCat tool. Wiesmann found that the results of applying NMT to legal texts were overall poor and detected 28 error categories. Nevertheless, the best results with regards to both “comprehensibility” and “correspondence between source and target text” were achieved by the translation of the law text, probably due to a weaker presence (with respect to the other genres evaluated) of legal language features which could present challenges to machine translation, e.g., syntactic complexity, formulaic and elliptical usage and abbreviations. Terminology errors detected in the translation of the law were mainly related to the inability of correctly translating legal terms pertaining to the South Tyrolean context (e.g., *Giunta provinciale* translated as *Provinzialrat* instead of *Landesregierung*). Finally, De Camillis (2021) carried out an automatic and manual evaluation of two South Tyrolean decrees machine-translated using eTranslation and ModernMT, the latter adapted with a relatively small parallel corpus of 22.500 segments. She found that ModernMT achieves a better performance, especially when adapted with in-domain data. In particular, MT output yielded by ModernMT has fairly good accuracy and fluency, whereas most of the errors observed are related to South Tyrolean legal-administrative terminology.

2.9 Terminology evaluation

A known limit of existing MT systems is that they struggle when handling specialized terminology (Koehn and Knowles 2017; Farajian *et al.* 2018; Dinu *et al.* 2019). Despite being one of the most crucial aspects in translation and machine translation, the evaluation of terminology accuracy in MT has been a less explored area in MT research (Haque *et al.* 2020: 150; Scansani 2020: 24). As seen in Section 2.6, a number of approaches have been proposed to achieve domain adaptation and to inject terminology in MT systems. Most of these works, however, only evaluate MT outputs using standard

metrics which measure overall performance (like BLEU), therefore failing to assess the accuracy in terminology translation (Scansani 2020: 24; Alam *et al.* 2021). The manual evaluation of terminology translation in MT research usually only takes place as part of overall MT quality assessments, for example when comparing SMT and NMT systems (among others, (Burchardt *et al.* 2017; Macketanz *et al.* 2017; Specia *et al.* 2017; Beyer *et al.* 2017)). This is often done by adopting the MQM error annotation framework (Lommel *et al.* 2014b), which also includes a category for terminology. According to Haque *et al.* (2020: 163), however, this coarse-grained error type is “an oversimplified attribute and does not consider various nuances of term translation errors” and is therefore deemed inadequate to assess terminology accuracy.

Only recently, a number of studies specifically addressed terminology evaluation in MT output. Scansani *et al.* (2017) carried out an automatic and manual assessment of terminology accuracy in PBMT for academic course catalogues translation, after training two engines with a parallel corpus and adding a bilingual termbase at training time. Automatic terminology assessment was carried out by extracting³⁴ the number of termbase entries appearing in the MT output and computing precision, recall and f-score.

Vintar (2018) carried out an automatic and human evaluation of the terminology translation of Google Translate NMT system compared to its earlier PBMT model for the Slovenian-English language pair in the domain of karstology. Automatic terminology evaluation was carried out by matching³⁵ term entries from an external termbase and counting the number of correct terms in the MT output. Manual terminology evaluation was based on three categories, namely *Correct*, *False* and *Omitted*.

Farajian *et al.* (2018) automatically evaluated term translation accuracy of an instance-based adapted MT system. They proposed an automatic terminology evaluation metric, *term hit rate* (THR), which computes the percentage of terms correctly translated by the MT system given a bilingual glossary. After counting

³⁴ Termbase entries were matched after lemmatising the hypothesis and reference sentences.

³⁵ Since Slovenian has a rich inflectional morphology, lemmatisation was carried out on both the term list and the evaluated sentences.

matched terms in the hypothesis sentences, the metric clips the counts of the matched terms by their frequency in the reference, in order to avoid overestimating systems which over-generate the same term.

Haque *et al.* (2019; 2020) carried out a fine-grained comparative evaluation of term translation accuracy in PBMT and NMT. They created a Hindi-English gold standard for terminology evaluation in the juridical domain, also taking “lexical and inflectional variations for a reference term” (2019: 4) into account. To assess term translation accuracy, they proposed a new terminology evaluation metric, *TermEval* (2019) and a fine-grained terminology error typology.³⁶ Although introducing term variants in terminology evaluation, however, as also Baldassarre (2021: 42) observed, the proposed taxonomy only takes morphological and syntactical features into account and is therefore not deemed adequate to conduct observations of terminological nature.

Scansani (2020) automatically evaluated term translation yielded by an instance-based adapted NMT system in the institutional academic domain (English > Italian). Term translation accuracy was assessed computing THR (Farajian *et al.* 2018) on the MAGMATic data set, a multi-domain academic gold standard with manual annotation of terminology (Scansani *et al.* 2019).

Exel *et al.* (2020) developed a terminology-constrained NMT model and automatically assessed term translation by reporting *term rates* (the percentage of correct terms generated in the MT output) and *variant term rates* (taking terminological variants from the reference termbase into account).

Baldassarre (2021) manually and automatically evaluated terminology translation in DeepL Pro enhanced with bilingual glossaries (English <> German, English <> French). Automatic term evaluation was carried out reporting THR scores (Farajian *et al.* 2018).

Scansani and Dugast (2021) compared term translation accuracy in four MT providers which allow user terminology integration, reporting TER and term *accuracy* scores (see below in this Section (Alam *et al.* 2021)).

³⁶ Details about categories and subcategories of the proposed taxonomy can be found in Haque *et al.* (Haque *et al.* 2020: 163–166).

Recently, automatic metrics which measure overall MT quality but specifically take terminology errors into account have been proposed. Dougal and Lonsdale (2020), for instance, proposed TREU (Terminology Recall Evaluation Understudy), which is comparable to BLEU but is more sensitive to injected terminology. Alam *et al.* (2021) proposed TERM, a terminology-biased modified version of TER, which penalizes errors that concern the terminology tokens more than other tokens. Terminology translation is assessed using a simple *accuracy* metric, computed as the number of matched terms in the MT output divided by the number of terms in source. TERM also checks whether terms are correctly placed in the hypothesis sentences.

2.10 Summing up

In this Chapter, the main developments in the field of machine translation have been reviewed, with particular focus on NMT, terminology, MT adaptation and evaluation and MT of legal texts. A number of studies has been carried out specifically on legal machine translation in the Italian-German language pair. Legal terminology specific to South Tyrol emerged as the main issue in today's commercial MT systems when translating into German, since they cannot be tuned according to diatopic parameters (Heiss and Soffritti 2018). Moreover, being an under-resourced language, resources in South Tyrolean German variety probably have little (if any) weight within the huge corpora used to train commercial generic MT systems (De Camillis and Contarino 2021). However, there has been no attempt to date to systematically adapt existing MT systems to the South Tyrolean specific legal terminology by means of parallel corpora and/or terminology.³⁷ In addition, to the best of our knowledge, no structured evaluation specifically addressing the translation of South Tyrolean legal terminology by MT has been carried out so far. In the present work, a first attempt of tailoring an MT system using South Tyrolean German resources will be made. Moreover, terminology accuracy

³⁷ The only attempt of MT domain adaptation (De Camillis 2021) has been carried out by means of a small in-domain corpus (approx. 22,500 sentence pairs), which has been deemed an insufficient amount of data to achieve a significant adaptation (2021: 291).

in MT output will be evaluated by proposing a fine-grained automatic evaluation pipeline, which will provide more detailed insights on the issue.

CHAPTER 3

3. METHODS

3.1 Introduction

In this Chapter, we will review the methodology adopted to conduct this dissertation project, providing details on each of its stages. Firstly, the aims and research questions of the study will be presented (3.2). We will then provide details about the methods applied to build the parallel corpus on which the study is based (3.3), to adapt an MT engine (3.4), estimate the overall quality of its output (3.5) and evaluate legal terminology accuracy (3.6).

3.2 Aims and research questions

As discussed in Chapter 2, MT has emerged as a useful resource to improve language accessibility and as a tool to support existing translation processes. This holds true particularly for international organisations and institutions of multilingual countries with high translation demands for documents in their official languages. In such contexts, MT systems are often trained or adapted by means of in-domain language resources in order to yield correct terminology and phraseology.

In South Tyrol, despite the strong needs for translated legal-administrative texts, MT has not been used or integrated in the institutional translation workflows yet. In general, to the best of our knowledge, there has not been a structured attempt of creating and evaluating an adapted MT system in the Italian-South Tyrolean German language pair. In order to be profitably used to translate legal-administrative texts, an MT system should be evaluated (either automatically or manually) and achieve scores that can be deemed “good enough” for gisting and post-editing purposes.¹ In this setting, moreover,

¹ Although not providing in-deep insights on the kind of errors generated by an MT engine, automatic metrics provide tangible information about MT performance and quality (see Sections 2.7 and 3.1). For a general interpretation of BLEU scores and their correlation to MT performance see, for example, the table provided in Google AutoML’s documentation (<https://cloud.google.com/translate/automl/docs/evaluate>, last accessed 12/11/2021).

a key aspect to be taken into account is local South Tyrolean legal-administrative terminology, which exhibits a series of peculiar features and is of central importance in institutional translation.

Although MT systems struggle when translating domain-specific terms, the evaluation of terminology has been a less explored field in MT research (see Section 2.9). In particular, only recently have a handful of studies specifically focused on evaluating terminology accuracy and only one taxonomy has been proposed for fine-grained terminology evaluation in MT output (Haque *et al.* 2020). However, the cited evaluation framework mainly focuses on morphological and syntactical features and is therefore not deemed adequate to evaluate terminology accuracy.

Against this background, the present dissertation aims at adapting and evaluating a neural adaptive MT system for the Italian-South Tyrolean German language pair in the legal domain, with particular focus on the evaluation of legal terminology accuracy in MT output. More specifically, the work aims at answering the following research questions:

- RQ1: *Can adaptive neural machine translation be profitably used to translate South Tyrolean administrative-legal texts?*
- RQ2: *To what extent does MT adaptation improve the translation of South Tyrolean legal-administrative terminology?*

To answer these questions, a parallel corpus of South Tyrolean legislation was created in order to achieve MT adaptation of an existing NMT system. Evaluation targeted both overall MT quality, based on automatic evaluation metrics (BLEU, chrF3, hLEPOR), and legal terminology accuracy, which was evaluated by automatically classifying correct and wrong legal terms within a fine-grained ad hoc taxonomy in order to analyse terminology translation improvements after MT adaptation. An overview of the project stages is outlined in Figure 4.

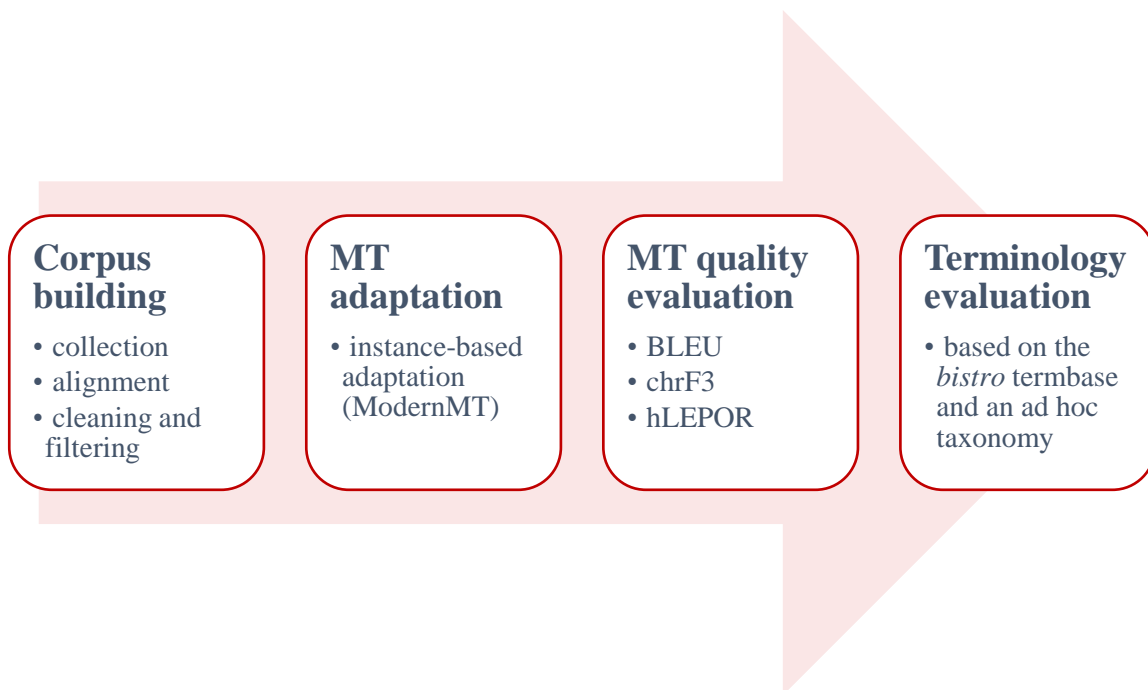


Figure 4: Overview of the stages of the present work.

3.3 Building the LEXB parallel corpus

LEXB is a bilingual parallel corpus of Italian and South Tyrolean German local and national legislation retrieved from the LexBrowser database, which features laws, decrees, resolutions, collective agreements and other national legal legislation of interest to South Tyrol, like the Italian Constitution, issued between 1946 and 08/02/2021). The corpus also contains a limited number of bilingual texts not published in the LexBrowser collection, namely 20 national laws and codes (Civil Code, Criminal Code) translated into German by the provincial Office for Language Issues.²

After having scraped pairs of URLs for each bitext,³ texts were collected and underwent a first stage of pre-processing and text filtering. Finally, texts have been sentence-aligned and the corpus has been cleaned at the sentence level and filtered in

² https://www.provincia.bz.it/politica-diritto-relazioni-estere/diritto/questioni-linguistiche/norme-statali-tradotte.asp?someforms_page=1&someforms_action=0 (last accessed: 31/10/2021).

³ A bitext is a pair of texts related to each other by means of *translational equivalence* (Tiedemann 2011: 7).

order to remove “bad” sentence pairs. Each step is described in greater detail in the following sections.

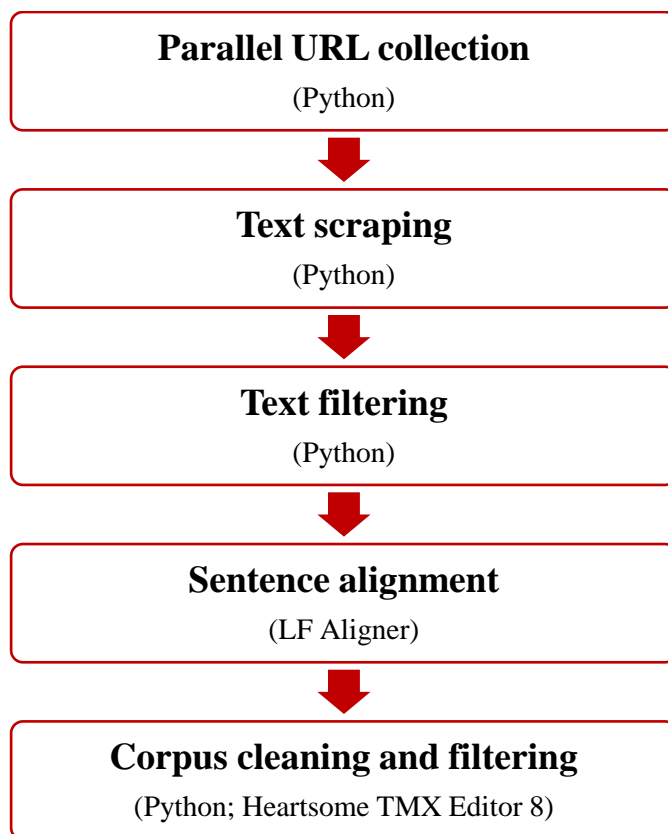


Figure 5: Overview on the LEXB corpus building steps and tools.

The final version of the corpus is in TMX format and contains approximately 175,000 sentence pairs and almost 9,500,000 tokens (see Table 1).

	LEXB (raw)	LEXB (final)
sentence pairs	280,210	174,468
tokens	12,486,758	9,479,569

Table 1: LEXB sentence pair and token counts, before and after cleaning and filtering operations.

3.3.1 Data collection and pre-processing

As seen in Section 1.5.2.1, texts on the LexBrowser can only be consulted monolingually, but a link to the translated version is available, which makes it possible to align texts at the document level. Texts in the LexBrowser collection are also

categorised according to the publication year, making it easy to collect all URLs of the texts hosted by the database by means of a simple web scraper.

Parallel URL collection and text extraction were carried out using a set of ad hoc Python scripts.⁴ As Makazhanov et. al (2018) observed, manual bitext extraction approaches based on in-house scripts yield cleaner bitexts and, subsequently, cleaner bisentences, significantly outperforming general, non-targeted, semi-automatic crawling approaches. As the purpose of this study was to retrieve high-quality parallel texts from a single website, it was deemed convenient to adopt a similar approach, by writing website-specific Python scrapers to address the tasks of parallel URL collection and bitext extraction.

During parallel URL collection, a first filtering stage was carried out to exclude texts which do not have a German translation but wrongly appear in the LexBrowser in both the Italian and German sections. Such preliminary filtering operation was based on blacklists of terms appearing in text titles.⁵ Afterwards, text collection was carried out by scraping the HTML source code of each of the collected URLs and extracting plain text without website boilerplate text. During the scraping stage, a further filtering operation was carried out to remove the remaining texts which appeared in the wrong language section in the LexBrowser. This was done by means of an automatic language identification algorithm. Plain texts were finally exported in .txt format. The final count of collected bitexts is 5007, including 20 texts translated into German by the provincial Office for Language Issues.

All scripts used for corpus creation and processing were written in Python 3.8. Python libraries used for scraping and text cleaning include *BeautifulSoup* and *urllib3* for URL and text scraping, *regex* for segment cleaning, *langid* for language detection and *xml* to handle XML special characters.

⁴ The scrapers for parallel URLs collection and texts extraction used to create the LEXB corpus are available at <https://github.com/antcont/LEXB>

⁵ For example, court decisions of the Regional Administrative Court (T.A.R.) and the Constitutional Court are rarely translated into German, but still wrongly appear in their Italian version in the German section of the LexBrowser.

3.3.2 Segmentation and sentence alignment

The following stage of text segmentation and sentence-level alignment was carried out using LF Aligner.⁶ LF Aligner is based on the *hunalign* algorithm (Varga *et al.* 2005) and has been chosen because it allows segmentation rules customisation (for instance, by adding domain-specific abbreviations) and batch automatic alignment.

Informal alignment accuracy assessment showed that automatic alignment accuracy yielded by LF Aligner was very high and, therefore, the output could be used “as-is” without the need for a manual revision of the entire corpus. Such high alignment quality is due to several factors. Firstly, South Tyrolean bilingual legal-administrative texts must have the same layout and typographical layout in both language versions (see Section 1.5.1). Moreover, institutional translators are required to preserve the same number of sentences when translating Italian laws into German (Chiocchetti *et al.* 2013a: 265), resulting in a “mirroring effect” between source and target texts (Woelk 2000: 216). As a consequence, segmentation discrepancies between source and target are very infrequent, as well as sentence inversions and reformulations, which therefore reduces the number of potential issues in automatic alignment. Secondly, a list of generic and legal-specific abbreviations was provided to the aligner’s integrated sentence segmentation tool. This made it possible to further reduce segmentation-related alignment errors, since abbreviations could otherwise be interpreted as end-of-sentence markers. Finally, a list of almost 20,000 bilingual legal term pairs extracted from the *bistro* terminological database was included into *hunalign*’s bilingual dictionary, which is leveraged by the algorithm during sentence alignment. This additional resource potentially contributed to further enhance alignment accuracy.

3.3.3 Corpus cleaning and filtering

As we have seen in Section 2.4, NMT systems have been proved to be by far more sensitive to noise in training data than SMT models, making the corpus cleaning and filtering steps a crucial task in both MT training and adaptation. In order to design and

⁶ <https://sourceforge.net/projects/aligner/>

achieve an effective corpus cleaning and filtering⁷ stage on the collected parallel data, the following operations were carried out:

- a) Possible types of noise in parallel corpora usually addressed when cleaning MT training data, as well as their influence on neural machine translation according to existing empirical research, were reviewed (see Section 2.4.1)
- b) A preliminary manual evaluation of the raw corpus was carried out, in order to identify the most frequently occurring noise types that need to be corrected or filtered out.
- c) Types of noise that needed to be tackled were defined, taking into account both their occurrence in the corpus and their possible influence on the quality of the MT system output
- d) An ad hoc parallel corpus cleaning pipeline was created, choosing the tools to be used for cleaning and filtering and developing scripts to carry out the necessary operations.

Preliminary evaluation allowed to gather important observations on the quality of the alignment and the main noise types displayed by the corpus. The observations can be summarised as follows:

- a) The overall quality of the automatic segmentation and alignment carried out by LF Aligner is very good (see Section 3.3.2).
- b) Several near-duplicate sentence pairs were identified, due to the character of legal texts. These near-duplicates are not expected to harm MT quality. Nevertheless, they can probably flaw experiments on MT customization, as many of them would probably be included both in the adaptation set and in the test set (see Section 3.4.1).
- c) Many segments were mainly composed of digits and punctuation.
- d) No major character encoding problems or corrupt characters are found in the corpus.

⁷ “Corpus cleaning” and “corpus filtering” are interchangeable terms usually used to indicate the task of removing bad sentence pairs from parallel corpora for MT training or MT adaptation. For the purposes of the present dissertation, however, “cleaning” refers to the *correction* of sentence-level noise, whereas “filtering” entails *removing* bad sentence pairs from the corpus.

- e) Some sentences contain excerpts of sentences in the other language. These may be harmful to the MT system and have therefore to be filtered out.
- f) Some sentences contain erroneously hyphenated words (e.g., “indica-zioni”, “Kompetenz-en”), which may be de-hyphenated during the cleaning stage.
- g) The most obvious and frequently occurring noise instances at the sentence level are list markers and superscript markers, which occur at the beginning or the end of segments, as well as substrings such as “Art. 1”, which are found at the beginning of many segments. Although being potentially useful anchors for the aligner, these occur with a very high frequency due to genre conventions and are deemed to be an element of noise in the LEXB corpus.

In light of the above observations, a pipeline was outlined for all corpus cleaning and filtering operations. The tools used for these purposes were:

- a) **Heartsome TMX Editor**:⁸ a powerful TMX editor which allows to carry out basic cleaning and filtering operations on translation memories. In the filtering pipeline of the LEXB corpus, Heartsome TMX Editor has only been adopted to filter out noise related to duplicate sentence pairs and inconsistencies in target, i.e., removing segments that have the same source text but a different translation.
- b) **tmx_cleaner.py**:⁹ an own in-house parallel corpus cleaning and filtering toolkit written in Python. It includes a set of ad hoc functions that allow to carry out all corpus cleaning and filtering operations on noise other than inconsistencies in target and simple deduplication.

In the following subsections, the single cleaning and filtering steps performed are presented in detail. Unless otherwise specified, the cleaning or filtering operation was carried out using the *tmx_cleaner.py* toolkit, which was developed specifically for the present work.

⁸ <https://github.com/heartsome/tmxeditor8>.

⁹ https://github.com/antcont/LEXB/blob/master/mt/tmx_cleaner.py.

3.3.3.1 Sentence-level cleaning

As observed during the preliminary evaluation, sentence-level noise (see Section 3.3.3, lit. g) is the most frequently occurring noise in the corpus. Removing such kind of corpus noise does not seem to be common practice when training MT systems and, to the best of our knowledge, there are no empirical studies on the influence of such segment-internal noise on the final MT quality. In the case of the LEXB corpus, however, since such noise typology affected most of the sentence pairs, it was deemed necessary to accurately clean up segments in order to obtain, as far as possible, only clean sentence pairs.

Noise at the beginning and the end of segments was cleaned up by means of a set of regular expressions. During this phase, the amount of useless sentence-internal text was reduced. Although we can only hypothesise that such sentence-level cleaning can bring slight benefits in terms of MT quality,¹⁰ it generated clean sentence pairs (which could be useful if the corpus is to be used as a translation memory) and lead to the identification of more duplicate sentence pairs in subsequent corpus filtering stages.

More than 275,000 single cleaning operations were carried out at the segment level in the LEXB parallel corpus. Details about the specific noise types removed during this stage are presented in Appendix A, Section 1.

3.3.3.2 Corpus filtering

Since corpus quality is fundamental to achieve good performance in neural machine translation, a number of filtering operations were carried out on the LEXB corpus. Noise in parallel corpora can have significant negative effects on NMT systems. Therefore, in designing our corpus filtering stage, we are more oriented towards precision than recall and tend therefore to be quite restrictive in choosing and tuning each of the operations to be carried out.¹¹ Details about the single filtering operations and parameters are presented in Appendix A, Section 2.

¹⁰ Verifying this hypothesis lays beyond the scope of the present work.

¹¹ Corpus filtering operations have been designed by taking into account the review carried out in Section 2.4.1.

The corpus filtering stage reduced the initial corpus size by approximately 38%. The largest filtering operation was sentence pair deduplication, which proves the repetitive nature of the text genres included in the LEXB corpus. The corpus filtering stage, including counts of single filtering operations, is summarised in Table 2.

<i>Filtering operation</i>	<i>Sentence pairs</i>	<i>%</i>
<i>Raw corpus</i>	280 210	100.00 %
Non-alphabetical/alphabetical ratio	3 069	1.10 %
Identical source-target	627	0.22 %
Highly similar source-target	1 963	0.70 %
Missing translation	0	0.00 %
Wrong language	93	0.03 %
Sentence length ratio	5 098	1.82 %
Long and short segments	42 181	15.05 %
Deduplication	49 137	17.54 %
Inconsistencies in target	3 574	1.28 %
<i>Cleaned and filtered corpus</i>	174 468	62.26 %

Table 2: Filtering operations carried out on the LEXB corpus: an overview.

3.4 Domain adaptation

3.4.1 Dataset splitting and near-duplicate processing

In machine learning experimental settings, the available datasets are usually divided into training set,¹² validation set¹³ and test set.¹⁴ The configuration and size of each set depend on the model, on the size of the dataset and on the task at hand. For the purposes of the present work, since we are carrying out real-time adaptation and not training an MT system from scratch, we split the data between adaptation set¹⁵ and test set. The test set consists of 2000 sentence pairs, which is a test set size widely adopted in the MT research community. Length of sentences included in the test set is between 10 and 20 tokens.

¹² The set of data used to train a machine learning algorithm (Zafar *et al.* 2018: 202).

¹³ Also called development set or dev set. The validation set serves the purpose of tuning the model's hyperparameters at intermediate stages of training (*ibid.*).

¹⁴ The set of unseen data that serves as gold standard for the final evaluation of a model (*ibid.*).

¹⁵ We call it *adaptation set* since we are not carrying out an actual MT *training*.

Before randomly selecting the test set, it is common practice to make sentence pairs in the dataset unique, i.e., to remove any duplicate sentence pair, which could appear in both the training/adaptation set and the test set and therefore flaw final results. However, as observed in the manual assessment of the parallel corpus (see Section 3.3.3), there are a number of highly similar sentence pairs (near-duplicates) that can “escape” simple sentence deduplication and therefore risk to flaw the results. Therefore, an additional advanced deduplication stage was carried out in order to tackle this issue. The operation of removing near-duplicates is not systematically carried out in MT research and NLP research¹⁶ and only rarely has it been pointed out as a potential problem.¹⁷ However, for datasets with highly repetitive sentence pairs due to genre-specific conventions, this step is undoubtedly crucial in order to get unbiased results from adaptation experiments.

The approach adopted for deduplication of near-duplicate sentence pairs is based on the “unique sentence pair filter” method applied in Tilde’s dataset filtering pipeline (Pinnis 2018), which consists in:

1. removing whitespaces and punctuation;
2. replacing digits with a placeholder;
3. lowercasing;
4. deduplicating based on the resulting normalised string representations.

Our modified approach also includes additional normalisation steps, which take into account genre-related features. In particular, variable legal-specific elements (text sections, text types, genre-specific numerals) which appear in similar sentence structures are also replaced with placeholders (see Table 3).

¹⁶ E-mail exchange with Benjamin Marie [27/09/2021].

¹⁷ Cfr. <https://logrusglobal.com/news/why-bleu-is-often-inflated.html>.

type	examples
text types	legge provinciale, L.P, decreto, Landesgesetz, L.G., Dekret, ...
text sections	articolo, art., titolo, comma, Artikel, Absatz, ...
genre-specific numerals	2ter, 4bis, 1duodecies, etc.
months	gennaio, febbraio, Jänner, Februar...

Table 3: Italian examples of elements replaced by placeholders for the purpose of near-duplicate sentence identification.

Overall, 6188 unique sentences were found to occur more than once in the corpus in the form of near-duplicates, with some of them occurring up to 3000 times as highly similar sentences. A total number of approximately 29000 occurrences of near-duplicates were identified in the corpus. Contrary to the approach adopted by Pinnis (2018), in our pipeline these sentence pairs are not removed, since such a reduction in the adaptation corpus size could influence the overall MT performance. However, we put near-duplicates in a blacklist, which kept them from being randomly included in the test set. Preliminary experiments carried out on a “biased” test set selected randomly and without this constraint showed that the presence of near duplicates can skew automatic scores by up to 15 BLEU points compared to an unbiased, deduplicated test set.

3.4.2 ModernMT

ModernMT (Bertoldi *et al.* 2018) is an adaptive neural machine translation system that allows users to integrate their own translation memories in order to adapt the MT output to the user’s terminology and style. The system also “learns” in real-time from human post-editing corrections, by immediately integrating corrected sentences in the pool of parallel data used to adapt the model for future translations. In addition, ModernMT achieves *document-level* adaptation, by generating a translation that is based on the content on the whole document, and not only the single sentence.¹⁸

More in detail, ModernMT adaptive system is based on the *instance-based adaptation* approach described by Farajian et al. (2017). Given an existing NMT

¹⁸ ModernMT approaches document-level adaptation by integrating a context vector when translating single sentences. The effectiveness of document-level adaptation has not been tested in the framework of the present dissertation, since testing and evaluation have been carried out at the sentence level.

baseline model, a pool of parallel data (including the in-domain adaptation corpus provided by the user as a translation memory) and a sentence to be translated, the method consists in retrieving from the parallel data a set of source-target sentence pairs in which the source is similar to the sentence to be translated. The parameters of the neural network model are then locally fine-tuned using the recalled sentence pairs. After translating the sentence, the adapted model is reset to the parameters of the original system. Apart from resulting in an overall quality improvement, this approach was also demonstrated to achieve a significant enhancement in terminology translation (Farajian et al. 2018).

ModernMT has been chosen to carry out MT adaptation in the present dissertation, since its adaptive algorithm, by fine-tuning the model on the fly, significantly boosts the accuracy of translated terminology and is less expensive, with regards to time and computational resources, than training a new system from scratch. Moreover, it allows to carry out MT adaptation in the Italian-German sentence pair, which is not offered by other commercial adaptive MT systems, like Microsoft Custom Translator or Google AutoML.

3.5 Overall quality evaluation

3.5.1 Evaluation metrics

After achieving domain adaptation by means of the LEXB corpus, overall quality improvement of the adapted ModernMT system is evaluated employing three automatic evaluation metrics (BLEU, chrF3, hLEPOR). MT quality is reported in terms of BLEU scores since, despite the limitations described in Section 2.7, it still is the *de facto* standard metric used in MT research and industry. chrF3 has been chosen because character-based metrics have shown a higher correlation to human judgements than word-based approaches. It is also suggested by Kocmi et al. (2021) as an alternative to state-of-the-art metrics based on multilingual language models for language pairs where no pre-trained model is available; in these scenarios, it proved to be the best-performing string-based metric. Finally, we report hLEPOR scores, since it is more precise with respect to word order, sentence length, precision and recall. BLEU and chrF3 are

computed using the SacreBLEU scripts (Post 2018), whereas hLEPOR is measured using a Python library recently made available by Logrus.¹⁹

3.5.2 Statistical significance testing

Improvements in quality scores between baseline and adapted system are tested for statistical significance using the *paired bootstrap resampling* test (Koehn 2004), a widely used significance test in MT research. Given a common test set machine translated using two different MT systems, the method consists in repeatedly creating “new virtual test sets by drawing sentences with replacement from the [test set] collection” (Koehn 2004). Corpus-level evaluation metric scores are then computed for both systems, and the “winning” system is noted. If one system outperforms the other in 95% of the total testing iterations, the conclusion can be drawn that it has a better performance in terms of quality with 95% statistical significance.²⁰ For the purpose of the present work, a paired bootstrap resampling test was carried out for each of the evaluation metrics adopted (BLEU, chrF3, hLEPOR) by resampling different test sets of 500 sentences for 1000 resampling iterations.

3.6 Terminology evaluation

As seen in Section 1.4, South Tyrolean German legal terminology features a number of peculiarities that differentiate it from legal terminologies pertaining to other German-speaking legal systems. Moreover, terminological variation is to a certain extent still a common phenomenon, resulting in the co-existence and co-occurrence of several terms designating a given concept. Some of these terms, moreover, may have been officially *standardised* by the Terminology Commission or be *recommended* for use in South Tyrol,²¹ whereas other terms could be outdated or be in use despite the standardisation

¹⁹ <https://pypi.org/project/hLepor/>

²⁰ See Koehn (2004) for more details about the test and how it is computed.

²¹ As we have seen in Section 1.4.1.4, after the Terminology Commission was discontinued, since 2015 a terminological harmonisation process has been carried out by the Institute of Applied Linguistics at Eurac Research and the provincial Office for Language Issues, as a result of which *recommended* terms are

of another term. For an example of terminological variation and overlapping of terms between German legal systems within a single concept, see Figure 6.

guida in stato di ebbrezza	IT	
guida sotto l'influenza dell'alcool	IT	
Alkohol am Steuer	Südtirol AT DE CH	
Trunkenheit am Steuer	Südtirol AT DE CH	<i>in Südtirol empfohlen</i>
Fahren in angetrunkenem Zustand	CH	
Trunkenheit im Verkehr	DE	
Trunkenheitsfahrt	DE	

Figure 6: Entry in the *bistro* database. The term “Trunkenheit am Steuer” is recommended for use in South Tyrol, with “Alkohol am Steuer” being an attested terminological variant. Finally, there are terms pertaining to other German-speaking legal systems only.

Considering the above-mentioned aspects, a simple classification between *correct* and *wrong* terms is not deemed accurate enough to carry out an informative evaluation of South Tyrolean German legal terminology translation. For example, a legal term yielded by an MT system could be the apparently correct translation of an Italian legal term but pertain to the Austrian, Swiss, or German legal system and be different from the actual correct term adopted in South Tyrol. Conversely, a term in use in South Tyrol is not necessarily the most correct solution, since it could be either an outdated term or a terminological variant used in place of a standardised or officially recommended term.

The aim of the proposed evaluation framework is therefore to carry out a fine-grained classification of legal terminology equivalents in the MT output of the ModernMT baseline and the domain-adapted system. This implies not only classifying terms as *correct* or *wrong*, but also automatically categorising legal terms within an *ad hoc* taxonomy according to:

- a) the *adequacy to the legal system*, i.e., whether the term used pertains to the South Tyrolean context specifically (contrary, for example, to terms that are used only in the German, Austrian or Swiss legal systems, or in European Union legislation);

flagged as such in the *bistro* database.

- b) the *term status*, i.e., whether the evaluated term is *standardised* by the Terminology Commission, *recommended* for use in South Tyrol, *obsolete* or attested in the *bistro* database as a correct term (although not standardised or recommended for use in the Province).²²

The classification is carried out automatically and is based on the *bistro* terminology database, which systematically contains important metadata at the term level, including information about the term status (standardised, recommended, obsolete) and about the legal system to which a term pertains (South Tyrol, Germany, Austria, Switzerland, etc.). Details and examples about the categories of the taxonomy are given in the following Section.

3.6.1 Evaluation taxonomy

The proposed fine-grained evaluation taxonomy contains three macro-categories for both correct and wrong classes. Two correct term categories, moreover, are further subdivided binarily. Therefore, the overall taxonomy contains 12 categories and, at the lowest level of granularity, a term can be classified as belonging to one of eight categories (see Figure 7 for a hierarchical overview of the taxonomy).

²² *Term status* is taken into account because terms standardised by the Terminology Commission are legally binding in documents published by public authorities (see Section 1.4.1.3) and recommended terms are the result of a terminological harmonisation process (see Section 1.4.1.4). Therefore, errors related to standardised or recommended terms are considered more severe than terms with no standardisation or official preference.

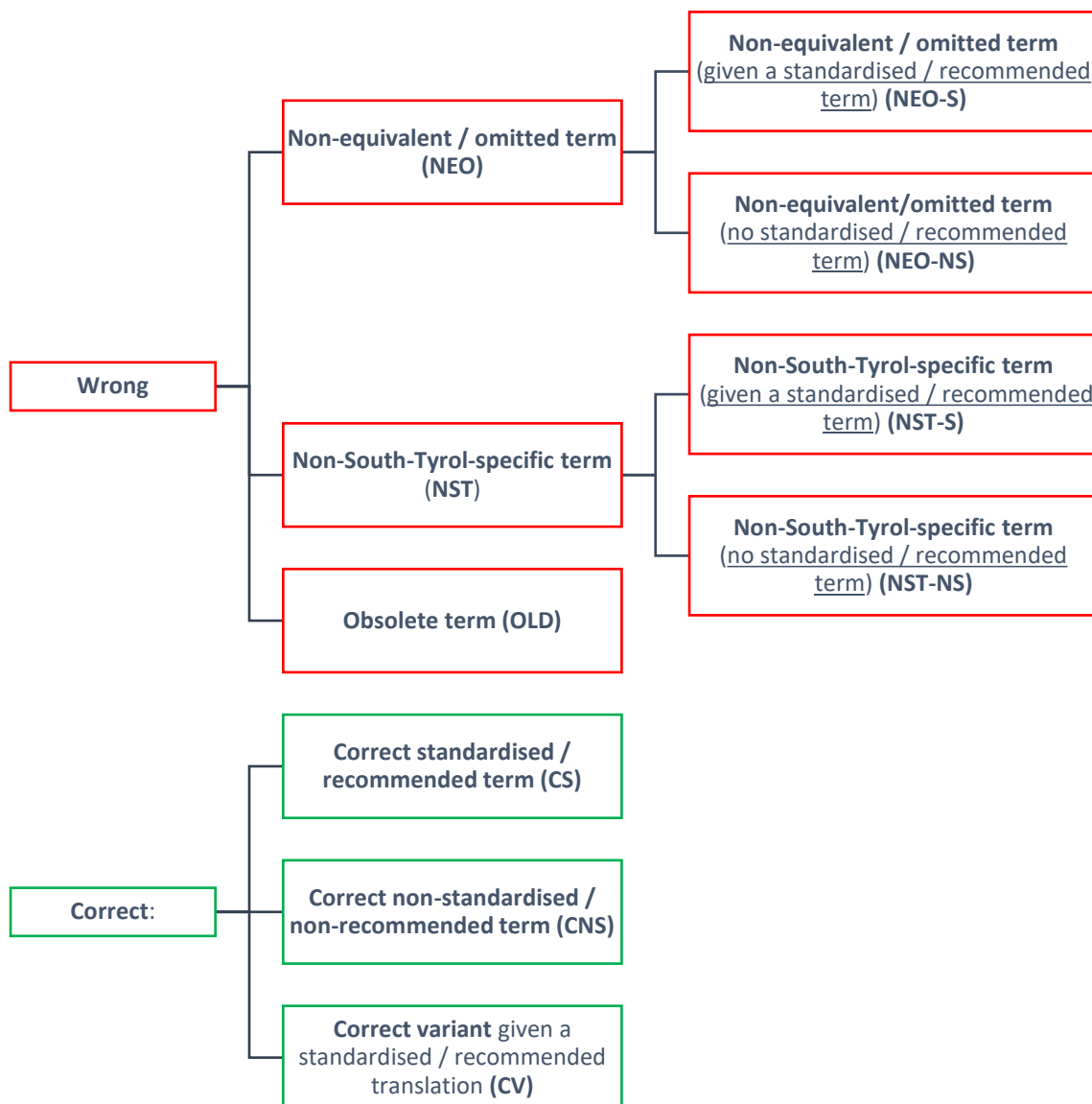


Figure 7: Overview of the proposed taxonomy for the evaluation of South Tyrolean legal terminology.

Wrong terms are classified into the following categories and subcategories:

1. **Non-equivalent/omitted term (NEO)**: The machine-translated sentence does not contain the term or any of its terminological variants. Subcategories include:

- a. **Non-equivalent/omitted term (given a standardised²³ or recommended²⁴ term) (NEO-S)**: An official translation of the Italian source term was standardised by the Terminology Commission or labelled as “recommended” for use in South Tyrol.
 - b. **Non-equivalent/omitted term (no standardised/recommended term) (NEO-NS)**: No official translation of the Italian source term was standardised by the Terminology Commission or labelled as “recommended” for use in South Tyrol.
2. **Non-South-Tyrol-specific term (NST)**: A term in the target sentence is not the correct South Tyrol-specific legal term. Although being terminologically equivalent, the term pertains to another German-speaking legal system. Subcategories include:
- a. **Non-South-Tyrol-specific term (given a standardised or recommended term) (NST-S)**: An official translation of the Italian source term was standardised by the Terminology Commission or labelled as “recommended” for use in South Tyrol.²⁵
 - b. **Non-South-Tyrol-specific term (no standardised / recommended term) (NST-NS)**: No official translation of the Italian source term was standardised by the Terminology Commission or labelled as “recommended” for use in South Tyrol.²⁶
3. **Obsolete term (OLD)**: a correct, South Tyrol-specific term was generated by the MT systems. The term, however, is outdated and therefore flagged as obsolete in the *bistro* terminology database.

²³ Terms standardised by the Terminology Commission and flagged as “*Südtirol genormt*” in the *bistro* termbase.

²⁴ Terms recommended for use in South Tyrol, which are flagged as “*in Südtirol empfohlen*” in the *bistro* termbase.

²⁵ Examples of this category are the terms *Fahren in angetrunkenem Zustand*, *Trunkenhiet im Verkehr* and *Trunkenheitsfahrt* (see Figure 6), since they pertain to another German-speaking legal system and a recommended translation of the term exists in South Tyrol.

²⁶ An example is the term *Exekutionshandlung* (see Figure 8), since it pertains to another German-speaking legal system. No South Tyrolean term, however, is standardised or recommended for this concept.

atto di esecuzione forzata	IT
Exekutionshandlung	AT
Zwangsvollstreckungshandlung	Südtirol DE

Figure 8: Entry from the bistro database (2).

Wrong term categories are sorted by decreasing level of error severity. The NEO category represents the most severe error in the taxonomy, since a non-equivalence/omission error is undoubtedly more critical than the use of a terminologically equivalent term, be it a term pertaining to another legal system (NST) or an outdated term (OLD). On a more granular level, error categories related to standardised or recommended terms (-S) are more severe than categories related to terms without an official preference (-NS).

Correct terms are automatically classified into the following categories:

1. **Correct standardised/recommended term (CS):** The source term is standardised by the Terminology Commission or is labelled as “recommended” for use in South Tyrol and the correct standardised/recommended term is used in the translated sentence.
2. **Correct non-standardised/non-recommended term (CNS):** The source term is neither standardised by the Terminology Commission nor labelled as “recommended” for use in South Tyrol, and the correct South Tyrol specific term is used in the translated sentence.²⁷
3. **Correct variant of a standardised/recommended term (CV):** The source term has a standardised/recommended translation for South Tyrol, but it was not used in the translated sentence. In place of the standardised term, however, a correct terminological variant attested in the *bistro* database was used.²⁸

²⁷ An example of term that would fall in this category is *Zwangsvollstreckungshandlung* (see Figure 8).

²⁸ Valid examples are, for instance, the terms *Quästur* and *Polizeipräsidium* used in place of *Polizeidirektion* (see Figure 9) or the term *Alkohol am Steuer* used in place of *Trunkenheit am Steuer* (see Figure 6).

questura	IT	
Polizeidirektion	Südtirol	Südtirol genormt
Polizeipräsidium	Südtirol	
Quästur	Südtirol	

Figure 9: Entry from the *bistro* database (3).

Correct term categories are sorted by decreasing level of correctness. The CS and CNS categories are equally correct, as the only difference consists in the status of the evaluated source term, i.e., whether a standardised/recommended term has been established or not. The CV category, instead, is considered slightly less correct, since the use of an attested variant term in place of the standardised/recommended term is undoubtedly a less appropriate choice.

3.6.2 Data pre-processing

Terminology classification is carried out automatically by retrieving relevant term-level information from the *bistro* database. To this purpose, the test set²⁹ and the terminological data contained in *bistro* underwent a series of pre-processing operations before being used to carry out automatic terminology classification and evaluation. Firstly, data from *bistro* was exported from SDL MultiTerm as an XML file, including only the relevant fields needed. Once exported, the XML file was parsed in order to build a Python data structure that could mirror the concept-oriented nature of the termbase and allow quick retrieval of terms and tags applied at the term level. At this stage, existing information contained in *bistro* was completed and converted to the tags defined in the proposed taxonomy (see Section 3.6.1). Moreover, entries with terms pertaining only to German-speaking legal systems other than South Tyrol were removed from the termbase. The final processed termbase used as a reference for the evaluation contains 9796 entries, each with an associated ID, and 31782 terms, each with tags about its geographical usage and status.

²⁹ The test set used for terminology evaluation is the same used for overall MT quality evaluation (see Section 3.5).

Finally, following the approach adopted by Vintar (2018), terms in the termbase and source, reference and target sentences are lemmatised using the TreeTagger (Schmid 1994), in order to achieve higher recall during automatic term matching.

3.6.3 Term matching and test set creation

The automatic terminology evaluation pipeline adopted in the present work consists of two stages: test set selection and term evaluation.³⁰ These steps are integrated into a common workflow, i.e., the entire process from term matching to the final annotation of evaluated terms is carried out for one test sentence at a time. Integrating test set selection and final evaluation without intermediate annotation of terms occurring in the source and reference sentences speeds up the evaluation process. Moreover, it makes it possible to quickly carry out terminology evaluation from scratch over any test set of source-reference-hypothesis sentence tuples without needing to pre-annotate them.

The first step serves the purpose of matching term pairs in the source and reference sentences of the test set and therefore define the evaluation benchmark, i.e., which target terms will be subsequently evaluated in the hypothesis sentence generated by the MT system. Term matching in our evaluation framework is carried out using spaCy's (Honnibal *et al.* 2020) PhraseMatcher,³¹ which efficiently matches large terminology lists over a given text input. spaCy's PhraseMatcher has been chosen because of its speed and, most importantly, because lookup units used for term matching (called *rules* in spaCy's documentation) allow the grouping of several search patterns (terms) under a common lookup unit. This makes it possible to maintain a concept-oriented approach by grouping terms from each terminological entry (concept) under one single lookup unit, together with their entry ID (see Figure 6).

³⁰ The algorithm for automatic terminology evaluation is made freely available under <https://github.com/antcont/LexTermEval/blob/main/LexTerm.py>.

³¹ <https://spacy.io/api/phrasematcher>.

```

matcher = PhraseMatcher(nlp.vocab)
matcher.add("24076", [nlp("delegazione legislativa")], [nlp("delega legislativa")])
matcher.add("24616", [nlp("fascicolo informatico")], [nlp("fascicolo processuale informatico")])

```

Figure 10: Examples of lookup units (rules) added to spaCy’s PhraseMatcher for term matching.

More in detail, for each sentence pair (source and reference) in the test set, the algorithm first looks for matches of the Italian terms contained in the termbase in the Italian source sentence. If no matches are found, the sentence pair is discarded from the test set, as it does not contain relevant legal terminology to be automatically evaluated. If more than one term match is found in the Italian sentence, matches are filtered greedily following a “first longest match” approach in order to avoid term annotation overlaps.³²

After filtering, for each matched Italian term the algorithm looks for matches of the South Tyrolean legal terms in the German reference sentence. If no matches are found, it splits German compound words using the CharSplit compound splitter³³ and runs the term matcher over the German sentence one more time, to allow matching of terms that are part of compound terms and could not be matched before splitting.³⁴ An example of term matched thanks to compound splitting is shown in Table 4.

	Italian	German
terms in entry	idoneità	Eignung
sentence	La disciplina sull' <u>idoneità al servizio</u> e sull'equo indennizzo è contenuta nell'allegato 5 al presente contratto.	In der Anlage 5 zum vorliegenden Vertrag ist die Regelung über die <u>Diensteignung</u> und die angemessene Entschädigung enthalten.

Table 4: Term matched in the German sentence thanks to compound splitting.

If no German terms are matched in the reference sentence, the sentence pair is discarded from the test set, since we are evaluating terms in MT output only if terms from the same *bistro* entry occur as a pair both in the source Italian sentence and in the German

³² As done by Farajian *et al.* (2018). For instance, given the sentence “All’articolo 4/bis del decreto del Presidente della Repubblica 28 marzo 1975, n. 474, sono apportate le seguenti modifiche” and the terms “decreto”, “Presidente della Repubblica” and “decreto del Presidente della Repubblica” in the reference termbase, the algorithm filters the matches greedily and only the translation of the term “decreto del Presidente della Repubblica” is evaluated.

³³ <https://github.com/dtugener/CharSplit>.

³⁴ spaCy’s PhraseMatcher matches strings at the token level, not at the subword level.

human reference sentence.³⁵ If more than one term match is found in the sentence, matches are filtered greedily. These filtered matches constitute the benchmark against which the MT hypothesis sentences will be evaluated. A flowchart representation of the pipeline designed for test set selection is given in Figure 11.

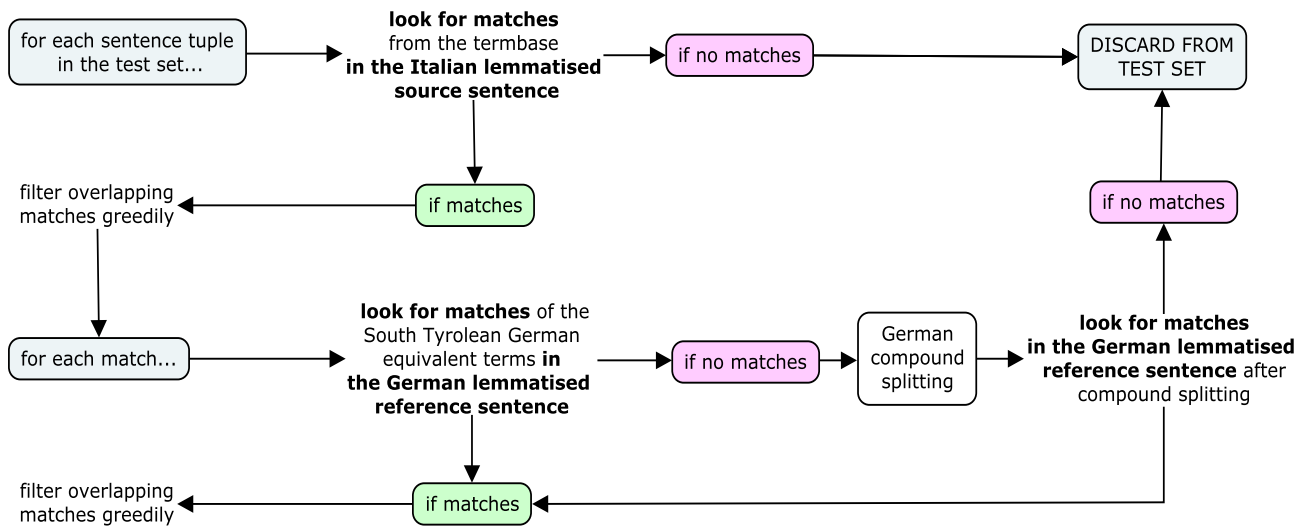


Figure 11: Flowchart of the pipeline for terminology matching (in the source and reference sentences) and test set selection.

3.6.4 Evaluation and annotation

After identifying term pairs in the source and reference sentences, the algorithm proceeds with the matching and evaluation of the legal terminology in the sentences generated by the MT system. More specifically, for each of the it-de term pairs matched in the preceding stage, the algorithm retrieves all German equivalent terms from the respective termbase entry and looks for matches of these terms in the hypothesis sentence (i.e., the sentence translated by the MT engine). If no matches are found in the first search, a second lookup is performed after splitting German compounds. If a match is found, the matched term is pointed to its base form in the terminology entry and the respective term classification tags are assigned. For each evaluated term, the algorithm

³⁵ If the respective equivalent is not found in the human reference translation, the source term may either be a homograph term referring to another concept (which may not be in the termbase) or not be a term at all.

tags the term as correct or wrong, and further applies a more granular classification according to the taxonomy described in Section 3.6.1. A flowchart representation of the final part of the evaluation pipeline is given in Figure 12.

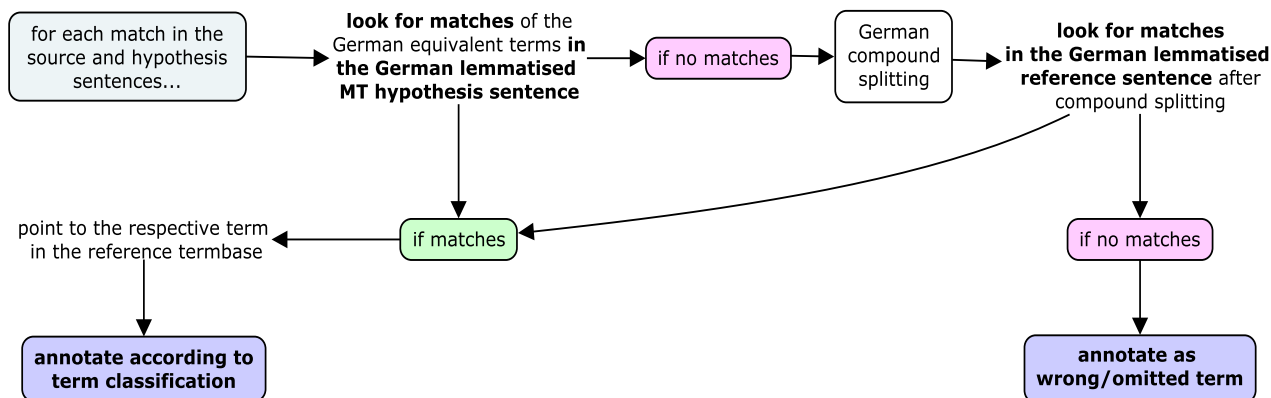


Figure 12: Flowchart of the pipeline for terminology matching (in the hypothesis sentence) and final term annotation.

At the end of the evaluation, two statistics are calculated: overall *term accuracy*, defined as the rate of correct terms out of all evaluated terms yielded by the MT system, and counts for each evaluation category. The annotated test data is also exported in tabular format: for each evaluation unit, the file contains the source-reference-hypothesis sentences, their lemmatised version, the concept entry ID, the terms contained in the terminological entry, the Italian matched term in the source sentence, the German term matched in the hypothesis sentence, the geographical usage of the evaluated term, a tag reporting if the term is correct, and the tag indicating the category in which the term is classified.

3.6.5 Advantages and limitations

The approach adopted to evaluate terminology translation in the present work is based on a completely automatic pipeline. On the one hand, a manual evaluation carried out by a translator or terminologist may be deemed more complete than an automatic assessment, since a human evaluator could evaluate additional terms that are not

matched by the automatic evaluation method.³⁶ However, a manual evaluation would also entail a number of disadvantages. Firstly, manual terminology evaluation can be complex, slow and therefore costly in terms of time and resources. Moreover, as any human-made evaluation activity, it is undoubtedly influenced by a certain degree of subjectivity. Like in the evaluation of overall MT quality, this can cause issues in inter-annotator agreement, both with regards to term selection³⁷ and term evaluation, especially if a given term has not been standardised, recommended or attested in any legal terminology database.

The proposed automatic evaluation method, on the contrary, features a number of advantages over a manual approach. Firstly, it is faster, as it allows to match and evaluate legal terminology over a large set of sentences within seconds. This translates into reduced costs in terms of time and resources. Secondly, it is reproducible, potentially more precise and objective, since it is based exclusively on a large, detailed and reliable legal terminology database for term matching and term equivalence classification. Unlike other automatic approaches, which do not usually allow fine-grained evaluation and do not uncover error severity,³⁸ the proposed method allows a more in-depth insight into the typologies of errors yielded by the MT system with respect to South Tyrolean legal terminology. Finally, unlike most terminology evaluation approaches in machine-translated text, which use the terms in the human-translated reference sentence as the point of reference for evaluation, the proposed approach is more concept-oriented and evaluates terminology taking terminological variants into account, too.

As with any fully automatic approach, the method adopted for the evaluation of terminology translation in the present work has some limitations. Firstly, it probably has a lower recall than manual evaluation, since term matching exclusively depends on the terms contained in the reference termbase and on the accuracy of the external libraries

³⁶ Terms may not be matched by the algorithm because they are not attested in the reference termbase or they were not processed correctly by the external libraries.

³⁷ The question of the “legal termhood” of a term is complicated by the open nature of legal language, as the separation between legal language and common language is not always clear-cut (Ralli 2009).

³⁸ Apart from a few exceptions, terminology evaluation in MT output is usually limited to a binary classification between correct and wrong terms (see Section 2.9).

used for text lemmatisation and compound splitting. Moreover, the method is not designed to evaluate terminology in machine translated sentences without relying on human reference translations, since term evaluation is only carried out if a term *pair* is found in both the source sentence and in the reference translation.

3.7 Summing up

In the present Chapter, the methodology adopted in this work has been presented. Firstly, details were provided about the aims and questions of the research, about the methods applied for building the LEXB parallel corpus, achieving domain adaptation and evaluating overall MT quality. Afterwards, the pipeline proposed for automatic terminology evaluation and the taxonomy adopted for term classification were presented. The proposed evaluation method can prove useful in several MT evaluation scenarios. In particular, it allows to instantly evaluate, based on a common test set, two or more MT systems with regards to South Tyrolean German legal terminology translation, by yielding a term accuracy score as well as a fine-grained classification of term equivalents. Moreover, the proposed method can also be applied as a pre-processing stage within a completely manual evaluation workflow, by serving as a data pre-annotation tool. In the present dissertation, it is used to evaluate, classify and compare term translation in MT output yielded by the ModernMT baseline system and the same system adapted by leveraging the cleaned and filtered version of the LEXB corpus.

CHAPTER 4

4. RESULTS

4.1 Introduction

In this Chapter the results of the study will be presented in detail and discussed. Firstly, results about the comparative quality evaluation between the generic and adapted MT system with regards to overall MT quality will be reported (4.2). Secondly, the outputs of the baseline and adapted MT systems will be evaluated in terms of legal terminology accuracy (4.3), following the evaluation framework described in Section 3.6. Finally, the obtained results will be discussed (4.4).

4.2 MT quality evaluation

In order to explore the potential of applying MT to the translation of South Tyrolean legal-administrative texts, the overall quality of the MT system adapted by leveraging the LEXB corpus has been evaluated by means of automatic metrics. Scores computed for the adapted system were compared with scores obtained by the ModernMT baseline system as well as by two state-of-the-art generic MT systems, Google Translate and DeepL.

Results are shown in Table 5. As can be noted, DeepL is the best performing system on our test set among the generic systems in terms of chrF3 and hLEPOR scores.¹ Results also show a particularly promising performance improvement achieved by the ModernMT domain-adapted system, which outperforms the ModernMT baseline system by +9 BLEU, +0.052 chrF3 and +0.048 hLEPOR points. Quality improvement in terms of both BLEU, chrF3 and hLEPOR is highly statistically significant ($p < 0.001$) according to a paired bootstrap resampling test carried out with 1000 resampling

¹ DeepL is outperformed by Google Translate in terms of BLEU score, but the difference is not statistically significant ($p > 0.05$) according to a paired bootstrap resampling test. DeepL's superiority on Google Translate, on the contrary, is statistically significant ($p < 0.05$) in terms of chrF3 and highly significant ($p < 0.001$) in terms of hLEPOR scores.

iterations on test sets of 500 sentences. These results show that leveraging the LEXB corpus as in-domain adaptation data has yielded a significant impact on translation quality, making the adapted MMT system the best performing system explored so far for the Italian-German language pair in the South Tyrolean legal-administrative domain. Following the guideline and scoring method proposed by Marie *et al.* (2021) for MT research papers relying on automatic metrics for the evaluation of translation quality (see Section 2.7), the automatic evaluation carried out in the present work can be considered “trustworthy”, since it does not rely exclusively on BLEU scores and it includes statistical significance testing to corroborate improvement claims.

	<i>BLEU</i>	<i>chrF3</i>	<i>hLEPOR</i>
<i>Google Translate</i>	27.61	0.529	0.637
<i>DeepL</i>	26.93	0.536	0.648
<i>ModernMT (baseline)</i>	25.73	0.517	0.626
<i>ModernMT (adapted)</i>	34.73	0.569	0.664

Table 5: Evaluation scores of generic MT systems vs. ModernMT domain-adapted system.

Despite uncovering the general extent of performance improvements related to domain adaptation, however, automatic metrics only give an approximate and overall view of translation quality. More specifically, they do not provide information on the specific translation features (grammar, fluency, accuracy, terminology, etc.) which improve after domain adaptation. In particular, with regards to legal terminology, which is a central aspect in the translation of legal-administrative documents, results in terms of automatic metrics cannot show the extent and significance of any improvements in terminology translation in the domain-adapted MT system. Therefore, in order to assess term accuracy, the output of the baseline and adapted ModernMT systems were comparatively analysed using the automatic terminology evaluation and classification method presented in Section 3.6.

4.3 Automatic evaluation of terminology translation

4.3.1 ModernMT baseline system

Automatic evaluation and classification of legal terminology in the output of ModernMT (baseline) and ModernMT (adapted) were carried out according to the methods and taxonomy presented in Section 3.6. In particular, based on the *bistro* reference termbase, legal terms were matched in the MT output and subsequently classified according to their term status and adequacy to the (Italian) legal system. Counts for each category as well as term accuracy, i.e., the rate of correct terms out of all evaluated terms, are reported for both systems and compared. Results concerning the output yielded by the ModernMT baseline system are shown in Table 6.

ModernMT baseline		
<i>category</i>	<i>tag</i>	<i>counts</i>
<i>correct</i>	<i>Correct standardised / recommended terms</i>	CS 891
	<i>Correct non-standardised / non-recommended terms</i>	CNS 1644
	<i>Correct variant terms given a standardised / recommended term</i>	CV 95
Total correct		2630
<i>wrong</i>	<i>Non-equivalent / omitted terms</i>	NEO 856
	⊥ <i>...given a standardised or recommended term</i>	⊥ NEO-S 539
	⊥ <i>... without standardised or recommended term</i>	⊥ NEO-NS 317
	<i>Non-South-Tyrol-specific terms</i>	NST 17
	⊥ <i>...given a standardised or recommended term</i>	⊥ NST-S 11
	⊥ <i>...no standardised or recommended term</i>	⊥ NST-NS 6
	<i>Obsolete terms</i>	OLD 0
Total wrong		873
<i>evaluated sentences</i>		1635
<i>evaluated terms</i>		3503
<i>term accuracy</i>		75.07 %

Table 6: Automatic terminology evaluation and classification of the MMT baseline system output.

As can be noted in Table 6, overall results are positive even without domain adaptation. The MMT baseline system managed to correctly translate 2630 legal terms out of 3503 evaluated terms in 1635 sentences,² therefore achieving 75.07% in term accuracy. The score is surprisingly positive considering that it refers to a generic, non-adapted system. Fine-grained classification makes it possible to have a more in-depth view into the type of correct (C) and wrong (W) legal terms yielded by the MT system. In particular, the highest number of errors (856) concerns the omission of terms and the use of non-equivalent terms (NEO error category, i.e., none of the terms from the respective reference terminological entry are matched in the evaluated sentence). More specifically, in the majority of such cases (539), the error concerns a term that is standardised or recommended for use in South Tyrol (NEO-S), which is considered the most severe error in our evaluation taxonomy. Many of such error instances are related to legal concepts that are highly specific to the Italian legal system and the Province of Bolzano (e.g., *legge provinciale*, *decreto del Presidente della Repubblica*, etc.). See Table 7 for an example of NEO-S error in the ModernMT baseline system.

	<i>sentence</i>	<i>C/W</i>	<i>category</i>
<i>src</i>	Agli effetti del presente regolamento , per « legge » si intende la legge provinciale 17 febbraio 2000 , n . 7.		
<i>hyp</i>	Im Sinne dieser Verordnung bezeichnet der Ausdruck " Gesetz » das Provinzgesetz Nr . 7 vom 17. Februar 2000.	W	NEO-S

Table 7: Example of NEO-S error in the MMT baseline system. The term “Provinzgesetz”, which is not attested as a correct term in the bistro database, was used in place of the standardised term “Landesgesetz”.

317 instances of non-equivalence/omission errors (NEO) refer to terms that are attested in the *bistro* database but are not officially harmonised (NEO-NS) and may therefore be considered slightly less severe errors. Like for NEO-S errors, many instances of NEO-NS errors are related to legal concepts that are highly specific to the South Tyrolean context (e.g., *giunta provinciale*, *Istituto provinciale di statistica*, etc.). See Table 8 for an example of NEO-NS error in the ModernMT baseline system.

² The test set coincides with the test set used for overall MT quality evaluation (see Section 4.2). Out of 2000 sentences, 365 did not contain any assessable legal term pair and were automatically discarded. Therefore, the test set for automatic terminology evaluation consists of 1635 sentence pairs.

	<i>sentence</i>	<i>C/W</i>	<i>category</i>
<i>src</i>	Detto importo viene aggiornato ogni cinque anni sulla base dell' indice nazionale dei prezzi al consumo redatto dall' Istituto provinciale di statistica .		
<i>hyp</i>	Dieser Betrag wird alle fünf Jahre auf der Grundlage des vom Statistischen Amt der Provinzen aufgestellten nationalen Verbraucherpreisindex aktualisiert .	W	NEO-NS

Table 8: Example of NEO-NS error in the MMT baseline system. The term “Statistische Amt der Provinzen”, is used in place of the correct term “Landesinstitut für Statistik”.

Only a reduced number of errors (17) concerns the use of terminologically equivalent terms pertaining to other German-speaking legal systems in place of the South-Tyrol-specific legal term (NST). See Table 9 for an example of NST error in the ModernMT baseline system.

	<i>sentence</i>	<i>C/W</i>	<i>category</i>
<i>src</i>	I beni immobili indicati alla lettera a) e b) del comma 1 del presente articolo si distinguono in :		
<i>hyp</i>	Die in Absatz 1 Buchstaben a) und b) genannten Grundstücke werden unterteilt in :	W	NST-S

Table 9: Instance of NST-S error in the MMT baseline system. The term “Grundstück”, which is attested in the bistro database as term pertaining to the German (Germany) legal system, is used in place of the standardised term “unbewegliche Sache”.

Finally, no errors related to the use of obsolete terms were identified by the automatic evaluation system. This is probably due to the under-representation of such terms in the reference termbase, since only 118 terms are labelled as obsolete out of 12550 total South Tyrolean German terms.

4.3.2 ModernMT adapted system

Automatic legal terminology evaluation was subsequently run over the output of the ModernMT system adapted with the LEXB in-domain parallel data. Results are shown in Table 10.

ModernMT adapted with LEXB

	<i>category</i>	<i>tag</i>	<i>counts</i>
<i>correct</i>	<i>Correct standardised / recommended terms</i>	CS	986
	<i>Correct non-standardised / non-recommended terms</i>	CNS	1664
	<i>Correct variant terms given a standardised / recommended term</i>	CV	96
	Total correct		2746
<i>wrong</i>	<i>Non-equivalent / omitted terms</i>	NEO	744
	⌞ <i>...given a standardised or recommended term</i>	⌞ NEO-S	446
	⌞ <i>... without standardised or recommended term</i>	⌞ NEO-NS	298
	<i>Non-South-Tyrol-specific terms</i>	NST	13
	⌞ <i>...given a standardised or recommended term</i>	⌞ NST-S	8
	⌞ <i>...no standardised or recommended term</i>	⌞ NST-NS	5
	<i>Obsolete terms</i>	OLD	0
	Total wrong		757
	<i>evaluated sentences</i>		
<i>evaluated terms</i>			3503
<i>term accuracy</i>			78.39 %

Table 10: Automatic terminology evaluation and classification of the MMT adapted system output.

Results show a substantial (although not particularly striking) improvement over the score and category counts obtained by the MMT baseline system (See Table 11 for a detailed comparison between the evaluated systems).

	<i>categories</i>	<i>MMT baseline</i>	<i>MMT adapted</i>	<i>% difference</i>
<i>correct</i>	CS	891	986	+ 10.66%
	CNS	1644	1664	+ 1.21%
	CV	95	96	+ 0.95%
	total	2630	2746	+ 4.41%
<i>wrong</i>	NEO	856	744	- 13.08%
	\sqsubseteq NEO-S	539	446	- 17.25%
	\sqsubseteq NEO-NS	317	298	- 5.99%
	NST	17	13	- 23.52%
	\sqsubseteq NST-S	11	8	- 27.27%
	\sqsubseteq NST-NS	6	5	- 16.6%
	OLD	0	0	
total	873	757	- 14.43%	
<i>evaluated terms</i>		3503	3503	
<i>term accuracy</i>		75.07%	78.39%	+ 3.31%

Table 11: Comparison of the evaluation results in the MMT baseline and adapted systems.

After leveraging the LEXB corpus as in-domain data, the ModernMT adapted system managed to correctly translate 2746 terms out of 3503 (+116 correct terms over the baseline system), therefore achieving 78.39% in legal term accuracy. Overall improvement in terms of total correct and wrong terms is statistically significant according to a McNemar test ($\chi^2 = 25.389$, $df = 1$, $p\text{-value} = 0.0000004687$). The most relevant improvements are observed with regards to the number of correct standardised/recommended terms translated correctly (CS) and the number of non-equivalence/omission errors (NEO), with 986 standardised/recommended terms translated correctly (vs. 891 by the baseline system) and 744 omitted/non-equivalent terms (vs. 856 by the baseline system). Improvements in other categories (CNS, CV, NST) are not significant (see Table 11 and Figure 13).



Figure 13: Comparison of correct and wrong term categories between the MMT baseline and the MMT adapted systems.

Improvements in terms of CS and NEO-S categories are of particular interest, since they are related to the use of standardised terms (which are legally binding in South Tyrol, see Section 1.4.1.2) and terms recommended for use in South Tyrol (see Section 1.4.1.3). More in detail, among the 539 NEO-S term errors yielded by the MMT baseline system, 190 have been translated correctly (CS) by the MMT adapted system. Examples are shown in Tables 12 and 13.

	<i>sentence</i>	<i>C/W</i>	<i>category</i>
<i>src</i>	All' articolo 4 / bis del decreto del Presidente della Repubblica 28 marzo 1975 , n . 474 , sono apportate le seguenti modifiche :		
<i>hyp</i>	Artikel 4 / bis des Präsidentialerlasses Nr . 474 vom 28. März 1975 wird wie folgt geändert :	W	NEO-S
<i>hyp2</i>	In Artikel 4 / b des Dekrets des Präsidenten der Republik vom 28. März 1975 , Nr . 474 , werden folgende Änderungen vorgenommen :	C	CS

Table 12: Example of NEO-S error in the MMT baseline system improved to a term of category CS after domain adaptation. The Italian term “decreto del Presidente della Repubblica” was translated by the adapted system using the correct standardised term “Dekret des Präsidenten der Republik”.

	<i>sentence</i>	<i>C/W</i>	<i>category</i>
<i>src</i>	L' inquadramento avviene sulla base della tabella di equiparazione prevista nell' allegato A 18) al presente decreto .		
<i>hyp</i>	Die Klassifizierung erfolgt auf der Grundlage der in Anhang A 18) dieses Dekrets vorgesehenen Ausgleichstabelle .	W	NEO-S
<i>hyp</i>	Die Einstufung erfolgt aufgrund der Gleichstellungstabelle gemäß Anlage A 18) zu diesem Dekret .	C	CS

Table 13: Example of NEO-S error in the MMT baseline system improved to a term of category CS after domain adaptation. The Italian term “inquadramento” was translated by the adapted system using the correct standardised term “Einstufung”.

A more reduced number of terms (19) categorised as NEO-S in the baseline system was translated correctly by the adapted system, but using a valid variant attested in the *bistro* database (CV) in place of the term standardised or recommended for South Tyrol (see Table 14).

	<i>sentence</i>	<i>C/W</i>	<i>category</i>
<i>src</i>	L' ammontare del sussidio concesso alle associazioni riconosciute non può superare il 60 per cento della spesa ammessa .		
<i>hyp</i>	Die den anerkannten Vereinigungen gewährte Subvention darf 60 % der zuschussfähigen Ausgaben nicht überschreiten .	W	NEO-S
<i>hyp2</i>	Der den anerkannten Vereinigungen gewährte Zuschuss darf 60 % der zuschussfähigen Ausgaben nicht überschreiten .	C	CV

Table 14: Instance of NEO-S error in the MMT baseline system improved to a term of category CV after domain adaptation. The Italian term “sussidio” was translated by the baseline system using the hypernym “Subvention” (NEO-S), whereas the adapted system yielded the attested variant “Zuschuss” in place of the standardised term “Beihilfe”.

Finally, 336 terms categorised as NEO-S in the MMT baseline system were translated with terms categorised as non-equivalent/omitted (NEO-S) even after domain adaptation. An example is shown in Table 15.

	<i>sentence</i>	<i>C/W</i>	<i>category</i>
<i>src</i>	Inoltre vengono applicati gli ambiti disciplinari del decreto ministeriale 23 febbraio 2016. n . 93.		
<i>hyp</i>	Darüber hinaus werden die Disziplinalggebiete des Ministerialerlasses Nr . 93 vom 23. Februar 2016 angewendet .	W	NEO-S
<i>hyp2</i>	Darüber hinaus gelten die Disziplinarbereiche des Ministerialerlasses Nr . 93 vom 23. Februar 2016.	W	NEO-S

Table 15: Instance of NEO-S error in both the MMT baseline and adapted systems. The Italian term “decreto ministeriale” was not translated with the standardised translation “Ministerialdekret”.

Apart from NEO-S, further term categories in the baseline system that improved to the CS category after domain adaptation include the NST-S³ and CV⁴ categories. Examples of terminology improvements from these categories are shown in Tables 16 and 17.

	<i>sentence</i>	<i>C/W</i>	<i>category</i>
<i>src</i>	Il recesso per giusta causa è regolato dall' articolo 2119 del codice civile .		
<i>hyp</i>	Der Rücktritt aus wichtigem Grund unterliegt Artikel 2119 des Bürgerlichen Gesetzbuches .	W	NST-S
<i>hyp2</i>	Der Rücktritt aus wichtigem Grund ist in Artikel 2119 des italienischen Zivilgesetzbuches geregelt .	C	CS

Table 16: Instance of NST-S error in the MMT baseline system improved to a term of category CS after domain adaptation. The Italian “codice civile” was translated by the MMT baseline system with “Bürgerliches Gesetzbuch”, term pertaining to the German (Germany) legal system, whereas it was correctly translated with the South Tyrolean standardised term “Zivilgesetzbuch” by the MMT adapted system.

	<i>sentence</i>	<i>C/W</i>	<i>category</i>
<i>src</i>	Nel contratto collettivo 8 marzo 2006 l’articolo 8 è così sostituito :		
<i>hyp</i>	Im Tarifvertrag vom 8. März 2006 erhält Artikel 8 folgende Fassung :	W	CV
<i>hyp2</i>	Im Kollektivvertrag vom 8. März 2006 erhält Artikel 8 folgende Fassung :	C	CS

Table 17: Instance of CV error in the MMT baseline system improved to a term of category CS after domain adaptation. Th Italian term “contratto collettivo” was translated by the MMT baseline system with “Tarifvertrag”, a variant of the standardised term “Kollektivvertrag” yielded by the adapted system.

The overall number of terms improved between the baseline and the adapted system is 340, including both W→C improvements and improvements within the *correct* and *wrong* macro-categories.⁵

Informal manual assessment of term evaluation instances carried out on a random sample basis showed that improvement of legal term translation is not always systematically consistent over different sentences, with several terms occurring multiple times in the test set and not being translated consistently by the MMT adapted system.

³ Non-South-Tyrol-specific terms given a standardised or recommended term.

⁴ Correct variant terms given a standardised or recommended term.

⁵ For instance, improvements from NEO to NST and from CV to CS are also included in the count.

The term *Giunta provinciale*, for instance, occurred 59 times in the test set and was translated by the MMT adapted system with the correct term *Landesregierung* only 16 times. The remaining instances were translated with the non-equivalent terms *Landesrat*, *Provinzialrat*⁶ and *Provinzrat*. See examples in Table 18.

	<i>sentence</i>	<i>C/W</i>	<i>category</i>
<i>src</i>	La Giunta provinciale può concedere un contributo fino all'80 per cento dei costi riconosciuti per altre strutture del Servizio antincendi .		
<i>hyp</i>	Der Provinzrat kann einen Beitrag von bis zu 80 Prozent der anerkannten Kosten für andere Feuerwehranlagen gewähren .	W	NEO-NS
<i>hyp2</i>	Die Landesregierung kann einen Beitrag bis zu 80 Prozent der anerkannten Kosten für andere Strukturen der Feuerwehr gewähren .	C	CNS
<i>src</i>	I limiti , le condizioni , le modalità di erogazione e il controllo delle prestazioni sono stabiliti dalla Giunta provinciale .		
<i>hyp</i>	Die Grenzen , Bedingungen , Methoden der Erbringung und Kontrolle der Dienstleistungen werden vom Provinzrat festgelegt .	W	NEO-NS
<i>hyp2</i>	Die Grenzen , die Bedingungen , die Art und Weise der Durchführung und die Kontrolle der Leistung werden vom Landesrat festgelegt .	W	NEO-NS
<i>src</i>	In sede di assegnazione del fondo di cui al comma 2 la Giunta provinciale può autorizzare di corrispondere degli acconti .		
<i>hyp</i>	Bei der Zuweisung des in Absatz 2 genannten Fonds kann der Provinzrat die Zahlung von Vorschüssen genehmigen .	W	NEO-NS
<i>hyp2</i>	Bei der Zuweisung des in Absatz 2 genannten Fonds kann der Provinzialrat die Zahlung von Vorschüssen genehmigen .	W	NEO-NS
<i>src</i>	La Giunta provinciale può riservare il 10 % del fondo sociale provinciale per il finanziamento di maggiori oneri imprevisti .		
<i>hyp</i>	Der Provinzrat kann 10 % des Provinzsozialfonds zur Finanzierung erhöhter unvorhergesehener Abgaben zurückstellen .	W	NEO-NS
<i>hyp2</i>	Der Provinzrat kann 10 % des Provinzsozialfonds für die Finanzierung höherer unvorhergesehener Kosten reservieren .	W	NEO-NS

Table 18: Instances of the translation of the term "Giunta provinciale" by the MMT baseline and adapted systems.

⁶ *Landesrat* and *Provinzialrat* are actually the equivalent terms for, respectively, *assessore provinciale* and *consigliere provinciale*.

Although to a lesser extent, the same phenomenon was also observed in the translation of the Italian terms *legge provinciale* and *Decreto del Presidente della Repubblica*. The reason behind such term inconsistency may be related to the instance-based adaptation approach adopted by the MMT system. Since the model adaptation is carried out on the fly by retrieving a set of sentences where the source is similar to the sentence to be translated (see Section 3.4.2), if few or no similar sentences are available in the pool of parallel data, domain adaptation may not be fully achieved at the sentence level and single terms may not be translated properly. This can be a drawback when the size of the available in-domain data is relatively small, as in the case of the LEXB corpus.

4.4 Discussion

Results described in Section 4.2 are particularly relevant to understand the potential of using adaptive NMT to translate legal-administrative texts in the Italian-South Tyrolean German language pair. Indeed, the automatic scores obtained by the adapted ModernMT system and the substantial improvement of 9 BLEU points over the performance of the baseline systems (DeepL, Google Translate, ModernMT baseline) show that using a system adapted by leveraging a corpus of in-domain data can be useful to profitably translate legal administrative texts. The adapted system performance improved considerably from the low scores achieved by the baseline systems (see Section 4.3.1) to a relatively good score of 35 BLEU points. Such results are statistically significant and the observed leap in performance is particularly promising considering the complexity of legal-administrative language, the relatively limited size of in-domain data (see Section 3.3) and the reduced similarity between adaptation set and training set, achieved by excluding near-duplicate sentence pairs from the test set (see Section 3.4.1).⁷

Moreover, quality improvement related to MT adaptation is also mirrored in terminology translation. This was uncovered by the fine-grained evaluation framework presented in Section 3.6. Instance-based adaptation yielded a significant improvement

⁷ As seen in Section 3.4.1, near-duplicates in the test set can inflate improvement statistics.

in the translation of legal-administrative terminology. Term accuracy improved by 3% between baseline and adapted system, with the latter being able to translate 116 more correct terms out the 3503 evaluated terms than the ModernMT generic system. In particular, the highest number of improved terms concerns standardised and recommended legal terms, which are highly crucial since they are legally binding or have been officially recommended for use in South Tyrol. These are very positive and promising results, since the correct translation of South-Tyrol-specific legal-administrative terms is the main issue in today's available NMT systems according to previous studies (Heiss and Soffritti 2018; Wiesmann 2019; De Camillis 2021).

On the other hand, terminology improvements observed are not exceptionally striking. Term accuracy did not reach top-level rates and critical problems, including terminology inconsistency, have been observed in the MT output (see Section 4.3.2). Despite achieving a statistically significant increase in term accuracy, instance-based adaptation with relatively limited adaptation data may therefore not be deemed the ideal approach to achieve a *systematic* enhancement⁸ in terminology translation in the considered language pair and domain.

⁸ An MT engine yielding correct terminology consistently should achieve a term accuracy of at least 95% (cfr., among others, Exel *et al.* 2020; Scansani and Dugast 2021).

CONCLUSIONS

The present work aimed at answering two research questions, i.e., “*Can adaptive neural machine translation be profitably used to translate South Tyrolean administrative-legal texts*” and “*To what extent does MT adaptation improve the translation of South Tyrolean legal-administrative terminology?*”. The questions were motivated by the high translation needs in the South Tyrolean public administrations, by the importance of legal-administrative terminology in institutional translation and by the challenges faced by existing machine translation systems when translating South Tyrolean legal-administrative terminology. At an applied level, the work also resulted in the creation of bilingual language resources in the Italian-South Tyrolean German language pair in the legal-administrative domain. To the best of my knowledge, the present study represents the first attempt of adapting a NMT system and automatically evaluating legal terminology accuracy in this language pair and domain.

After collecting, aligning, cleaning and filtering a relatively large number of bilingual legal-administrative texts to create the LEXB parallel corpus (3.3), the resource was used to perform domain adaptation of an MT engine through ModernMT (3.4). To answer the research questions, the evaluation phase tackled both overall quality, by means of automatic metrics (3.5), and legal terminology accuracy, by means of an ad hoc automatic evaluation method (3.6). More specifically, automatic legal terminology evaluation consisted in using an external reference terminology database (*bistro*) to match legal terms in translated sentences and automatically categorise them as *correct* and *wrong* terms, as well as, on a finer level of granularity, within 8 categories according to the term’s *adequacy to the legal system* and the *term status* (3.6.1).

Results presented in Sections 4.2 and 4.3 showed that instance-based adaptation achieved by leveraging an in-domain corpus of decrees, laws, resolutions, and agreements can bring about a substantial improvement in MT performance when translating legal-administrative texts in the Italian-South Tyrolean German language pair. The scores achieved by the ModernMT system domain-adapted by leveraging the LEXB corpus outperformed the generic systems (DeepL, Google Translate, ModernMT baseline) by 9 BLEU points, a highly statistically significant improvement in

performance, yielding a relatively good score of 35 BLEU. The results obtained are particularly promising considering the limited size of adaptation data (3.3) and the relatively reduced similarity between adaptation set and test set (3.4.1).

The quality improvement obtained with domain adaptation was also noticeable at the terminological level. The automatic terminology evaluation approach adopted (3.6) allowed to uncover and analyse terminology accuracy improvements both on a quantitative and qualitative level. The domain-adapted ModernMT system correctly translated 2746 out of 3503 legal terms of the test set (term accuracy: 78.39%), outperforming the generic ModernMT system by 3.31%. The most substantial improvements were observed with regards to standardised/recommended legal terms, which are highly crucial since they are legally binding, or have been defined as recommended for use in the legal-administrative texts produced by the local public administrations. Although statistically significant, the observed improvements in terminology accuracy have been deemed promising but not fully satisfactory if the aim is to achieve a systematic enhancement in legal terminology translation.

Automatic scores regarding overall MT quality and terminology accuracy therefore suggest that adaptive NMT could be profitably adopted to translate legal-administrative texts in the Italian-South Tyrolean German language pair. As far as legal terminology translation is concerned, however, despite the significant improvements in term translation accuracy, automatic terminology evaluation has shown that instance-based adaptation does not achieve a systematic enhancement in legal terminology translation.

The present work contributes to the field of MT research in the German-Italian language pair and in the legal-administrative domain in several ways. Firstly, the LEXB bilingual corpus has been created for the Italian-South Tyrolean German language pair in the legal-administrative domain to train and test MT systems. Evaluations of MT output quality carried out for the generic systems and the adapted system created by leveraging the LEXB corpus have pointed to overall MT performance improvements. Secondly, an ad hoc pipeline for automatic legal terminology evaluation for the Italian-South Tyrolean German language pair has been proposed. The method provides detailed quantitative and qualitative insights into legal terminology translation in MT output, by automatically matching legal terms and categorising them within a fine-grained

taxonomy according to term-level metadata from an external reference termbase. The scripts used for the creation of the LEXB corpus and for automatic terminology evaluation have been made freely available.⁹

The main limitations of the study are related to the size of the collected parallel data and the automatic nature of the adopted terminology evaluation approach. Despite containing all legal-administrative texts published in the LexBrowser database, the LEXB corpus has a relatively limited size for MT training purposes (175k sentence pairs, 9.5m tokens). Moreover, it contains texts issued since 1946, with several texts published before the establishment of the Terminology Commission in 1988. Therefore, we cannot rule out the possibility that sentences in the corpus may contain outdated or incorrect terminology, hence potentially influencing the effectiveness of domain adaptation on terminology accuracy. As for the method proposed for automatic terminology evaluation, limitations include that its accuracy is heavily contingent on external dependencies, on the set of terms in the reference termbase¹⁰ and on the presence of human reference translations.¹¹ Because of time constraints, moreover, it was not possible to systematically evaluate the accuracy of the evaluation method itself and the impact of such limitations in terms of precision and recall. This can be seen as the main limitation of the study.

Future work may overcome some of the above-mentioned limitations. The LEXB corpus, for instance, could be extended by collecting and aligning a larger number of parallel texts from the local administrations or by means of data augmentation techniques (e.g., back-translation), in order to further improve MT performance. Further research might also include investigating the effectiveness of other approaches for domain adaptation and terminology enhancement, including direct terminology integration, as well as benchmarking new systems using the proposed automatic

⁹ Scripts for URL and text scraping, corpus cleaning, corpus filtering, evaluation of overall MT quality and statistical significance testing are available at <https://github.com/antcont/LEXB>. Scripts for automatic terminology evaluation are available at <https://github.com/antcont/LexTermEval>.

¹⁰ This can lead to a lower recall than a manual assessment carried out by human evaluators (see Section 3.6.5).

¹¹ Term evaluation is only carried out if a term pair is found in both source and reference sentences (see Section 3.6.5).

terminology evaluation method and comparing results with those achieved in the present work. Future studies may also aim at exploiting the proposed automatic terminology evaluation method as a pre-annotation tool in order to manually integrate terminology annotations and to create a gold standard for automatic legal terminology evaluation in MT output. Finally, further experimental investigations are needed to assess the feasibility of successfully integrating MT in the translation workflows of the South Tyrolean public administrations, for example by carrying out a thorough manual evaluation as well as experiments in real-world scenarios on different legal-administrative text types and subdomains.

REFERENCES

- Aggarwal, Charu C. (2018) *Neural Networks and Deep Learning: A Textbook*, New York: Springer.
- Alam, Md Mahfuz ibn et al. (2021) ‘On the Evaluation of Machine Translation for Terminology Consistency’, *ArXiv:2106.11891 [Cs]*.
- Alber, Elisabeth and Palermo, Francesco (2012) ‘Creating, Studying and Experimenting with Bilingual Law in South Tyrol: Lost in Interpretation?’, in X. Arzoz (ed.). *Bilingual Higher Education in the Legal Context. Group Rights, State Policies and Globalisation*, Leiden: Nijhoff, 287–309.
- Alcock, Anthony (2001) *The South Tyrol Autonomy. A Short Introduction*, Londonderry/Bolzano: Autonomous Province of Bolzano.
- ALPAC (1966) *Language and Machines - Computers in Translation and Linguistics. A Report by the Automatic Language Processing Advisory Committee*, Washington, D.C.: National Academy of Sciences, National Research Council.
- Ammon, Ulrich (1995) *Die Deutsche Sprache in Deutschland, Österreich Und Der Schweiz. Das Problem Der Nationalen Varietäten*, Berlin/Boston: De Gruyter.
- Ammon, Ulrich et al. (2004) *Variantenwörterbuch des Deutschen: die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*, Berlin: De Gruyter.
- Armentano Oller, Carme et al. (2007) *Apertium, una plataforma de código abierto para el desarrollo de sistemas de traducción automática*, Cádiz: Servicio de Publicaciones de la Universidad de Cádiz.
- Ash, Tom, Francis, Remi and Williams, Will (2018) ‘The Speechmatics Parallel Corpus Filtering System for WMT18’, in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, October 2018, Belgium, Brussels: Association for Computational Linguistics, 853–859.

- ASTAT (2012) ‘Censimento Della Popolazione 2011’, *ASTAT Info* 38.
- ASTAT (2020) *Südtirol in Zahlen/Alto Adige in Cifre*, Bolzano: Landesinstitut für Statistik / Istituto provinciale di statistica.
- Bahdanau, Dzmitry, Cho, Kyunghyun and Bengio, Yoshua (2014) ‘Neural Machine Translation by Jointly Learning to Align and Translate’, *ArXiv:1409.0473 [Cs, Stat]*.
- Baker, Mona (1995) ‘Corpora in Translation Studies. An Overview and Suggestions for Future Research’, *Target* 7(2): 223–243.
- Baldassarre, Valentina Vincenza (2021) *Intégration de la terminologie en traduction automatique : Le cas de la Confédération suisse avec DeepL Pro*. Master thesis. Université de Genève.
- Banerjee, Satanjeev and 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*, June 2005, Ann Arbor, Michigan: Association for Computational Linguistics, 65–72.
- Barrault, Loïc et al. (2020) ‘Findings of the 2020 Conference on Machine Translation (WMT20)’, in *Proceedings of the 5th Conference on Machine Translation (WMT)*, 2020, 1–55.
- Bentivogli, Luisa, Bisazza, Arianna, Cettolo, Mauro and Federico, Marcello (2016) ‘Neural versus Phrase-Based Machine Translation Quality: A Case Study’, in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, November 2016, Austin, Texas: Association for Computational Linguistics, 257–267.
- Bertoldi, Nicola, Caroselli, Davide and Federico, Marcello (2018a) ‘The ModernMT Project’, in *Proceedings of The 21st Annual Conference Of the European Association for Machine Translation 28–30 May 2018* Universitat d’Alacant Alacant, Spain, 345.

- Bertoldi, Nicola, Caroselli, Davide and Federico, Marcello (2018b) ‘The ModernMT Project’, in *Proceedings of the 21st Annual Conference of the European Association for Machine Translation: 28-30 May 2018*, 2018, Alacant: European Association for Machine Translation, 345.
- Beyer, Anne, Macketanz, Vivien, Burchardt, Aljoscha and Williams, Philip (2017) ‘Can Out-of-the-Box NMT Beat a Domain-Trained Moses on Technical Data’, in *EAMT 2017: The 20th Annual Conference of the European Association for Machine Translation*, 31 May 2017.
- Brown, Peter F., Della Pietra, Stephen A., Della Pietra, Vincent J. and Mercer, Robert L. (1993) ‘The Mathematics of Statistical Machine Translation: Parameter Estimation’, *Computational Linguistics* 19(2): 263–311.
- Burchardt, Aljoscha et al. (2017) ‘A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines’, *The Prague Bulletin of Mathematical Linguistics* 108(1): 159.
- Castilho, Sheila et al. (2017) ‘Is Neural Machine Translation the New State of the Art?’, *The Prague Bulletin of Mathematical Linguistics* 108: 109–120.
- Castilho, Sheila, Doherty, Stephen, Gaspari, Federico and Moorkens, Joss (2018) ‘Approaches to Human and Machine Translation Quality Assessment’, in J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty (eds). *Translation Quality Assessment: From Principles to Practice*, Cham: Springer International Publishing, 9–38.
- Chan, Sin-Wai (2015) *The Routledge Encyclopedia of Translation Technology*, London: Routledge.
- Chatterjee, Rajen et al. (2017) ‘Guiding Neural Machine Translation Decoding with External Knowledge’, in *Proceedings of the Second Conference on Machine Translation*, September 2017, Copenhagen, Denmark: Association for Computational Linguistics, 157–168.

- Chen, Boxing, Kuhn, Roland, Foster, George, Cherry, Colin and Huang, Fei (2016) ‘Bilingual Methods for Adaptive Training Data Selection for Machine Translation’, in S. Green and L. Schwartz (eds). *Proceedings of the 12th Conference of The Association for Machine Translation in the Americas*, 2016, 93–106.
- Chen, Guanhua, Chen, Yun, Wang, Yong and Li, Victor O. K. (2020) *Lexical-Constraint-Aware Neural Machine Translation via Data Augmentation*, in 9 July 2020, 3587–3593.
- Chen, Pinzhen, Bogoychev, Nikolay and Germann, Ulrich (2020) ‘Character Mapping and Ad-Hoc Adaptation: Edinburgh’s IWSLT 2020 Open Domain Translation System’, in *Proceedings of the 17th International Conference on Spoken Language Translation*, July 2020, Online: Association for Computational Linguistics, 122–129.
- Chiocchetti, Elena (2019a) ‘Implementing Language Rights in South Tyrol: Developing Legal Terminology for the Public Administration and the Judiciary’, in I. Horvath (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)*, 2019, Budapest: ELTE Eötvös Kiadó, 103–117.
- Chiocchetti, Elena (2019b) ‘Legal Comparison in Terminology Work: Developing the South Tyrolean German Legal Language’, in S. Szoták (ed.). *Diszciplínák Találkozója: Nyelvi Közvetítés a XXI. Században*, Budapest: OFFI, 175–185.
- Chiocchetti, Elena (2019c) ‘Terminology Work in South Tyrol: New Approaches, New Termbase, New Contents’, *Terminologija* 26: 6–23.
- Chiocchetti, Elena, Kranebitter, Klara, Ralli, Natascia and Isabella, Stanizzi (2013a) ‘Deutsch ist nicht gleich Deutsch. Eine terminologische Analyse zu den Besonderheiten der deutschen Rechtssprache in Südtirol’, in M. M. Brambilla, J. Gerdes, and C. Messina (eds). *Diatopische Variation in der deutschen Rechtssprache*, Berlin: Frank & Timme, 253–285.

- Chiocchetti, Elena, Kranebitter, Klara, Ralli, Natascia and Stanizzi, Isabella (2019) ‘25 Jahre Bozner Methode: Terminologearbeit in Südtirol’, in P. Drewer and D. Pulitano (eds). *Terminologie: Epochen, Schwerpunkte, Umsetzungen: Zum 25-Jährigen Bestehen Des Rats Für Deutschsprachige Terminologie*, Berlin: Springer Vieweg, 175–191.
- Chiocchetti, Elena and Ralli, Natascia (2013) *Guidelines for Collaborative Legal / Administrative Terminology Work*, Bolzano: Eurac Research.
- Chiocchetti, Elena and Ralli, Natascia (2016) ‘Ein Begriff, zwei Sprachen, unterschiedliche (Rechts)Kulturen’, in P. Drewer, F. Mayer, and K.-D. Schmitz (eds). *Terminologie und Kultur. Akten des Symposions, Mannheim, 3.–5. März., 2016*, München: Deutscher Terminologie-Tag e.V, 103–112.
- Chiocchetti, Elena, Ralli, Natascia, Lusicky, Vesna and Wissik, Tanja (2013b) ‘Spanning Bridges between Theory and Practice: Terminology Workflow in the Legal and Administrative Domain’, *Comparative Legilinguistics. International Journal of Legal Communication* 16: 7–22.
- Chiocchetti, Elena, Ralli, Natascia and Stanizzi, Isabella (2013c) ‘When Language Becomes Law: The Methodology and Criteria Adopted by the South Tyrolean Terminology Commission for the Standardisation of German and Italian Translation Equivalents’, *Linguistica* 53(2): 9–23.
- Chiocchetti, Elena, Ralli, Natascia and Stanizzi, Isabella (2017) ‘From DIY Translations to Official Standardization and Back Again? 50 Years of Experience with Italian and German Legal Terminology Work in South Tyrol’, in P. Faini (ed.). *Terminological Approaches in the European Context*, Cambridge: Cambridge Scholars Publishing, 254–270.
- Chiocchetti, Elena and Stanizzi, Isabella (2010) ‘Die Beschlüsse Der Südtiroler Terminologiekommission: Problematiken Bei Der Normung von Rechtstermini’, in C. Heine and J. Engberg (eds). *Reconceptualizing LSP. Online Proceedings of the XVII European LSP Symposium 2009*, Aarhus: Aarhus Business School/Aarhus University.

- Cho, Kyunghyun et al. (2014) ‘Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation’, *ArXiv:1406.1078 [Cs, Stat]*.
- Chu, Chenhui, Dabre, Raj and Kurohashi, Sadao (2017) ‘An Empirical Comparison of Simple Domain Adaptation Methods for Neural Machine Translation’, *ArXiv:1701.03214 [Cs]*.
- Chu, Chenhui and Wang, Rui (2018) ‘A Survey of Domain Adaptation for Neural Machine Translation’, in *Proceedings of the 27th International Conference on Computational Linguistics*, August 2018, Santa Fe, New Mexico, USA: Association for Computational Linguistics, 1304–1319.
- Coluccia, Stefania (2000) ‘Il Linguaggio Giuridico in Alto Adige’, in A. Pasinato (ed.). *Heimat - Identità Regionali Nel Processo Storico*, Roma: Donzelli, 378–388.
- Conneau, Alexis and Lample, Guillaume (2019) ‘Cross-Lingual Language Model Pretraining’, in *Advances in Neural Information Processing Systems*, 2019, Vancouver, Canada: Curran Associates, Inc., 7057–7067.
- Cronin, Michael (2010) ‘The Translation Crowd’, *Tradumàtica* 8: 1–7.
- De Camillis, Flavia (2017) *Relazione Finale Di Progetto. I Processi Traduttivi Nell’amministrazione Provinciale*, Bolzano: Eurac Research.
- 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. Università di Bologna.
- De Camillis, Flavia and Contarino, Antonio Giovanni (2021) *Adapting machine translation for under-resourced languages: a first attempt for institutional German in South Tyrol*, PaCor 2021, Vitoria-Gasteiz, 23-25/06/2021, <https://hdl.handle.net/10863/17781> [Conference presentation]
- Dillinger, Mike and Lommel, Arle (2004) *LISA Best Practice Guide: Implementing Machine Translation*, Geneva: Localization Industry Standards Association.

- Dinu, Georgiana, Mathur, Prashant, Federico, Marcello and Al-Onaizan, Yaser (2019) ‘Training Neural Machine Translation to Apply Terminology Constraints’, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, July 2019, Florence, Italy: Association for Computational Linguistics, 3063–3068.
- Doddington, George (2002) ‘Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics’, in *Proceedings of the Second International Conference on Human Language Technology Research*, 24 March 2002, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 138–145.
- Dougal, Duane K. and Lonsdale, Deryle (2020) ‘Improving NMT Quality Using Terminology Injection’, in *Proceedings of the 12th Language Resources and Evaluation Conference*, May 2020, Marseille, France: European Language Resources Association, 4820–4827.
- Exel, Miriam, Buschbeck, Bianka, Brandt, Lauritz and Doneva, Simona (2020) ‘Terminology-Constrained Neural Machine Translation at SAP’, in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, November 2020, Lisboa, Portugal: European Association for Machine Translation, 271–280.
- Farajian, M. Amin, Bertoldi, Nicola, Negri, Matteo, Turchi, Marco and 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*, 2018, Alicante: European Association for Machine Translation, 149–158.
- Farajian, Mohammad Amin, Turchi, Marco, Negri, Matteo and Federico, Marcello (2017) ‘Multi-Domain Neural Machine Translation through Unsupervised Adaptation’, in *Proceedings of the Second Conference on Machine Translation*, September 2017, Copenhagen: Association for Computational Linguistics, 127–137.

- Farzindar, Atefeh and Lapalme, Guy (2009) ‘Machine Translation of Legal Information and Its Evaluation’, in *Proceedings of the 22nd Canadian Conference on Artificial Intelligence: Advances in Artificial Intelligence*, 15 May 2009, Berlin, Heidelberg: Springer-Verlag, 64–73.
- Garg, Ankush and Agarwal, Mayank (2019) ‘Machine Translation: A Literature Review’, *CoRR* abs/1901.01122.
- Gruber, Alfons (2000) *Geschichte Südtirols: Streifzüge Durch Das 20. Jahrhundert*, Bolzano: Athesia.
- Gupta, Rohit, Lambert, Patrik, Patel, Raj and 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*, August 2019, Dublin, Ireland: European Association for Machine Translation, 142–148.
- Gurney, Kevin (1997) *An Introduction to Neural Networks*, London: UCL.
- Han, Aaron L. F. et al. (2013) ‘Language-Independent Model for Machine Translation Evaluation with Reinforced Factors’, in *Machine Translation Summit XIV*, 2013, 215–222.
- Han, Aaron L. F., Wong, Derek F. and Chao, Lidia S. (2012) ‘LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors’, in *Proceedings of COLING 2012: Posters*, December 2012, Mumbai, India: The COLING 2012 Organizing Committee, 441–450.
- Han, Lifeng, Jones, Gareth J. F. and Smeaton, Alan F. (2021a) ‘Translation Quality Assessment: A Brief Survey on Manual and Automatic Methods’, *ArXiv:2105.03311 [Cs]*.
- Han, Lifeng, Sorokina, Irina, Erofeev, Gleb and Gladkoff, Serge (2021b) ‘CushLEPOR: Customised HLEPOR Metric Using LABSE Distilled Knowledge Model to Improve Agreement with Human Judgements’, *ArXiv:2108.09484 [Cs]*.

- Hangya, Viktor and Fraser, Alexander (2018) ‘An Unsupervised System for Parallel Corpus Filtering’, in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, October 2018, Belgium, Brussels: Association for Computational Linguistics, 882–887.
- Haque, Rejwanul, Hasanuzzaman, Mohammed and Way, Andy (2019) ‘TermEval: An Automatic Metric for Evaluating Terminology Translation in MT’, in *Proceedings of CICLing 2019, the 20th International Conference on Computational Linguistics and Intelligent Text Processing*, 2019, La Rochelle, France.
- Haque, Rejwanul, Hasanuzzaman, Mohammed and Way, Andy (2020) ‘Analysing Terminology Translation Errors in Statistical and Neural Machine Translation’, *Machine Translation* 34: 149–195.
- Heinisch, Barbara and Lušicky, Vesna (2020) ‘The Austrian Language Resource Portal for the Use and Provision of Language Resources in a Language Variety by Public Administration – a Showcase for Collaboration between Public Administration and a University’, in *Proceedings of the 1st Workshop on Language Technologies for Government and Public Administration (LT4Gov)*, May 2020, Marseille, France: European Language Resources Association, 28–31.
- Heiss, Christine Antonie and 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)*.
- Hildebrand, Almut Silja, Eck, Matthias, Vogel, Stephan and Waibel, Alex (2005) ‘Adaptation of the Translation Model for Statistical Machine Translation Based on Information Retrieval’, in *Proceedings of the 10th EAMT Conference: Practical Applications of Machine Translation*, 30 May 2005, Budapest, Hungary: European Association for Machine Translation.

- Hokamp, Chris and Liu, Qun (2017) ‘Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search’, in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2017, Vancouver, Canada: Association for Computational Linguistics, 1535–1546.
- Honnibal, Matthew, Montani, Ines, Van Landeghem, Sofie and Boyd, Adriane (2020) *SpaCy: Industrial-Strength Natural Language Processing in Python*, Zenodo.
- Hutchins, W. John (1995) ‘Machine Translation: A Brief History’, in *Concise History of the Language Sciences. From the Sumerians to the Cognitivists*, Elsevier, 431–445.
- Hutchins, W. John (2010) ‘Machine Translation: A Concise History’, *Journal of Translation Studies* 13: 29–70.
- Hutchins, W. John (2011) ‘Recent Applications of Machine Translation’, in K. Malmkjær and K. Windle (eds). *The Oxford Handbook of Translation Studies*, Oxford: Oxford University Press.
- Hutchins, William John and Somers, Harold (1992) *An Introduction to Machine Translation*, London: Academic Press.
- International Organization for Standardization (2007) *ISO 860:2007 Terminology Work — Harmonization of Concepts and Terms*, Geneva: International Organization for Standardization.
- International Organization for Standardization (2012) *ISO 10241-2:2012 Terminological Entries in Standards — Part 2: Adoption of Standardized Terminological Entries*, Geneva: International Organization for Standardization.
- International Organization for Standardization (2017) *ISO 18587:2017 Translation Services — Post-Editing of Machine Translation Output — Requirements*, Geneva: International Organization for Standardization.

- Jalili Sabet, Masoud, Negri, Matteo, Turchi, Marco, C. de Souza, José G. and Federico, Marcello (2016) ‘TMop: A Tool for Unsupervised Translation Memory Cleaning’, in *Proceedings of ACL-2016 System Demonstrations*, August 2016, Berlin, Germany: Association for Computational Linguistics, 49–54.
- Junczys-Dowmunt, Marcin, Dwojak, Tomasz and Hoang, Hieu (2016) ‘Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions’, *ArXiv:1610.01108 [Cs]*.
- Junczys-Dowmunt, Marcin, Pinero, Blanca and Ziemski, Michal (2015) *SMT at the International Maritime Organization: Experiences with Combining In-house Corpora with Out-of-domain Corpora*, in 11 May 2015.
- Kalchbrenner, Nal and Blunsom, Phil (2013) ‘Recurrent Convolutional Neural Networks for Discourse Compositionality’, *ArXiv:1306.3584 [Cs]*.
- Khayrallah, Huda and 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*, July 2018, Melbourne, Australia: Association for Computational Linguistics, 74–83.
- Killman, Jeffrey (2014) ‘Vocabulary Accuracy of Statistical Machine Translation in the Legal Context’, in *Third Workshop on Post-Editing Technology and Practice*, 2014, 85.
- Kit, Chunyu and Wong, Tak Ming (2008) ‘Comparative Evaluation of Online Machine Translation Systems with Legal Texts’, *Law Library Journal* 100(2): 299–322.
- Kocmi, Tom et al. (2021) ‘To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation’, *ArXiv:2107.10821 [Cs]*.
- Koehn, Philipp (2004) ‘Statistical Significance Tests for Machine Translation Evaluation’, in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, July 2004, Barcelona, Spain: Association for Computational Linguistics, 388–395.

- Koehn, Philipp (2009) *Statistical Machine Translation*, New York: Cambridge University Press.
- Koehn, Philipp et al. (2020) ‘Findings of the WMT 2020 Shared Task on Parallel Corpus Filtering and Alignment’, in *Proceedings of the Fifth Conference on Machine Translation*, November 2020, Online: Association for Computational Linguistics, 726–742.
- Koehn, Philipp (2020) *Neural Machine Translation*, Cambridge: Cambridge University Press.
- Koehn, Philipp, Guzmán, Francisco, Chaudhary, Vishrav and Pino, Juan (2019) ‘Findings of the WMT 2019 Shared Task on Parallel Corpus Filtering for Low-Resource Conditions’, in *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, August 2019, Florence, Italy: Association for Computational Linguistics, 54–72.
- Koehn, Philipp, Khayrallah, Huda, Heafield, Kenneth and Forcada, Mikel L. (2018) ‘Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering’, in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, October 2018, Belgium, Brussels: Association for Computational Linguistics, 726–739.
- Koehn, Philipp and Knowles, Rebecca (2017a) ‘Six Challenges for Neural Machine Translation’, in *Proceedings of the First Workshop on Neural Machine Translation*, August 2017, Vancouver: Association for Computational Linguistics, 28–39.
- Koehn, Philipp and Knowles, Rebecca (2017b) ‘Six Challenges for Neural Machine Translation’, in *Proceedings of the First Workshop on Neural Machine Translation*, August 2017, Vancouver: Association for Computational Linguistics, 28–39.
- Koehn, Philipp, Och, Franz J. and Marcu, Daniel (2003) ‘Statistical Phrase-Based Translation’, in *Proceedings of the 2003 Human Language Technology*

- Conference of the North American Chapter of the Association for Computational Linguistics*, 2003, 127–133.
- Koskinen, Kaisa (2008) *Translating Institutions. An Ethnographic Study of EU Translation*, Manchester: St. Jerome.
- Koskinen, Kaisa (2011) ‘Institutional Translation’, in Y. Gambier and L. van Doorslaer (eds). *Handbook of Translation Studies*, Amsterdam: John Benjamins.
- Koskinen, Kaisa (2014) ‘Institutional Translation: The Art of Government by Translation’, *Perspectives (Gerontological Nursing Association (Canada))* 22(4): 479–492.
- Kurfalı, Murathan and 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)*, August 2019, Florence, Italy: Association for Computational Linguistics, 277–281.
- Lample, Guillaume, Ott, Myle, Conneau, Alexis, Denoyer, Ludovic and Ranzato, Marc’Aurelio (2018) ‘Phrase-Based & Neural Unsupervised Machine Translation’, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.
- Lardilleux, Adrien and Lepage, Yves (2017) ‘CHARCUT: Human-Targeted Character-Based MT Evaluation with Loose Differences’, in *Proceedings of IWSLT 2017*, December 2017, Tokyo, Japan.
- Läubli, Samuel, Sennrich, Rico and Volk, Martin (2018) ‘Has Machine Translation Achieved Human Parity? A Case for Document-Level Evaluation’, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, October 2018, Brussels, Belgium: Association for Computational Linguistics, 4791–4796.
- Li, Xiaoqing, Zhang, Jiajun and Zong, Chengqing (2016) ‘One Sentence One Model for Neural Machine Translation’, *ArXiv:1609.06490 [Cs]*.

- LISA (2005) *TMX 1.4b Specification*, 2005, GALA Global.
- Liu, Qun and Zhang, Xiaojun (2015) ‘Machine Translation’, in S.-W. Chan (ed.). *The Routledge Encyclopedia of Translation Technology*, London: Routledge, 105–119.
- Liu, Yang and Zhang, Min (2015) ‘Statistical Machine Translation’, in S.-W. Chan (ed.). *The Routledge Encyclopedia of Translation Technology*, London: Routledge, 201–212.
- Lommel, Arle et al. (2014a) ‘Using a New Analytic Measure for the Annotation and Analysis of MT Errors on Real Data’, in *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, 16 June 2014, Dubrovnik, Croatia: European Association for Machine Translation, 165–172.
- Lommel, Arle (2018) ‘Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies’, in J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty (eds). *Translation Quality Assessment: From Principles to Practice*, Cham: Springer International Publishing, 109–127.
- Lommel, Arle, Uszkoreit, Hans and Burchardt, Aljoscha (2014b) ‘Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics’, *Tradumàtica* 12: 455–463.
- Lu, Jun, Lv, Xiaoyu, Shi, Yangbin and Chen, Boxing (2018) ‘Alibaba Submission to the WMT18 Parallel Corpus Filtering Task’, in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, October 2018, Belgium, Brussels: Association for Computational Linguistics, 917–922.
- Luong, Minh-Thang and Manning, Christopher D. (2015) ‘Stanford Neural Machine Translation Systems for Spoken Language Domain’, in *International Workshop on Spoken Language Translation*, 2015, Da Nang, Vietnam.
- Macketanz, Vivien, Avramidis, Eleftherios, Burchardt, Aljoscha, Helcl, Jindrich and Srivastava, Ankit (2017) ‘Machine Translation: Phrase-Based, Rule-Based and

- Neural Approaches with Linguistic Evaluation’, *Cybernetics and Information Technologies* 17(2): 28–43.
- Mai, Katja (2016) *Use of MT@EC by translators in the European Commission*, 2nd ELRC Conference, Bruxelles.
- Makazhanov, Aibek, Myrzakhmetov, Bagdat and Assylbekov, Zhenisbek (2018) ‘Manual vs Automatic Bitext Extraction’, in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, May 2018, Miyazaki, Japan: European Language Resources Association (ELRA).
- Marie, Benjamin and Fujita, Atsushi (2018) ‘Unsupervised Neural Machine Translation Initialized by Unsupervised Statistical Machine Translation’, *ArXiv:1810.12703 [Cs]*.
- Marie, Benjamin, Fujita, Atsushi and Rubino, Raphael (2021) ‘Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers’, in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, August 2021, Online: Association for Computational Linguistics, 7297–7306.
- Mathur, Nitika, Baldwin, Timothy and Cohn, Trevor (2020) ‘Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics’, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 2020, Online: Association for Computational Linguistics, 4984–4997.
- Mayer, Felix (2000) ‘Terminographie Im Recht: Probleme Und Grenzen Der Bozner Methode’, in D. Veronesi (ed.). *Linguistica Giuridica Italiana e Tedesca*, Padova: Unipress.
- McEnery, Tony and Hardie, Andrew (2012) *Corpus Linguistics: Method, Theory and Practice*, Cambridge: Cambridge University Press.

- McEnery, Tony, Xiao, Richard and Tono, Yukio (2006) *Corpus-Based Language Studies: An Advanced Resource Book*, New York: Routledge.
- Melby, Alan K. and Wright, Sue Ellen (2015) ‘Translation Memory’, in S.-W. Chan (ed.). *The Routledge Encyclopedia of Translation Technology*, London: Routledge, 662–677.
- Miceli Barone, Antonio Valerio, Haddow, Barry, Germann, Ulrich and Sennrich, Rico (2017) ‘Regularization Techniques for Fine-Tuning in Neural Machine Translation’, in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, September 2017, Copenhagen, Denmark: Association for Computational Linguistics, 1489–1494.
- Michon, Elise, Crego, Josep and Senellart, Jean (2020a) ‘Integrating Domain Terminology into Neural Machine Translation’, in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, Barcelona, Spain (Online): International Committee on Computational Linguistics, 3925–3937.
- Michon, Elise, Crego, Josep and Senellart, Jean (2020b) ‘Integrating Domain Terminology into Neural Machine Translation’, in *Proceedings of the 28th International Conference on Computational Linguistics*, December 2020, Barcelona, Spain (Online): International Committee on Computational Linguistics, 3925–3937.
- Mikolov, Tomas, Chen, Kai, Corrado, Greg and Dean, Jeffrey (2013) ‘Efficient Estimation of Word Representations in Vector Space’, *ArXiv:1301.3781 [Cs]*.
- Mileto, Fiorenza (2019) ‘Post-Editing and Legal Translation’, *H2D/Revista de Humanidades Digitais* 1(1).
- Miyata, Rei et al. (2016) ‘MuTUAL: A Controlled Authoring Support System Enabling Contextual Machine Translation’, in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, December 2016, Osaka, Japan: The COLING 2016 Organizing Committee, 35–39.

- Müller, Berndt, Reinhardt, Joachim and Strickland, Michael T. (1995) *Neural Networks: An Introduction*, 2nd Ed., Berlin Heidelberg: Springer-Verlag.
- Nagao, Makoto (1984) ‘A Framework of a Mechanical Translation between Japanese and English by Analogy Principle’, in *Proceedings of the International NATO Symposium on Artificial and Human Intelligence*, 1 October 1984, USA: Elsevier North-Holland, Inc., 173–180.
- Nießen, Sonja, Och, Franz Josef, Leusch, Gregor and 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)*, May 2000, Athens, Greece: European Language Resources Association (ELRA).
- Nurminen, Mary and Koponen, Maarit (2020) ‘Machine Translation and Fair Access to Information’, *Translation Spaces* 9(1): 150–169.
- O’Brien, Sharon, Choudhury, Rahzeb, van der Meer, Jaap and Aranberri, Nora (2011) *Dynamic Quality Evaluation Framework*, De Rijp: TAUS.
- Okpor, Margaret Dumebi (2014) ‘Machine Translation Approaches: Issues and Challenges’, *IJCSI International Journal of Computer Science Issues* 11(5): 159–165.
- Papineni, Kishore, Roukos, Salim, Ward, Todd and 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*, July 2002, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 311–318.
- Peterlini, Oskar (2000) *Autonomie Und Minderheitenschutz in Südtirol Und Trentino: Überblick Über Land Und Geschichte, Recht Und Politik*, Bolzano: Regionalrat der Autonomen Region Trentino-Südtirol.

- Peterlini, Oskar (2011) ‘L’autonomia – Strumento Di Pace’, in F. M. Provenzano (ed.). *Federalismo, Devolution, Secessione – Ritorno al Futuro*, Cosenza: Pellegrini, 293–317.
- Pierce, George (2018) *Introducing Translational Studies*, Waltham Abbey: ED-Tech Press.
- Pinnis, Mārcis (2015) *Terminology Integration in Statistical Machine Translation*. Doctoral thesis. University of Latvia.
- Pinnis, Mārcis (2018) ‘Tilde’s Parallel Corpus Filtering Methods for WMT 2018’, in *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, October 2018, Belgium, Brussels: Association for Computational Linguistics, 939–945.
- Popović, Maja (2015) ‘ChrF: Character n-Gram F-Score for Automatic MT Evaluation’, in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, September 2015, Lisbon, Portugal: Association for Computational Linguistics, 392–395.
- Popović, Maja (2018) ‘Error Classification and Analysis for Machine Translation Quality Assessment’, in J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty (eds). *Translation Quality Assessment: From Principles 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*, October 2018, Brussels, Belgium: Association for Computational Linguistics, 186–191.
- Post, Matt and Vilar, David (2018) ‘Fast Lexically Constrained Decoding with Dynamic Beam Allocation for Neural Machine Translation’, in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, June 2018, New Orleans, Louisiana: Association for Computational Linguistics, 1314–1324.

- Pouliquen, Bruno (2017) *WIPO Translate: Patent Neural Machine Translation publicly available in 10 languages*, Patent and Scientific Literature Translation workshop (MT Summit 2017), Nagoja, Japan.
- Pouliquen, Bruno, Elizalde, Cecilia, Junczys-Dowmunt, Marcin, Mazenc, Christophe and Garcia-Verdugo, Jose (2013) 'Large-scale multiple language translation accelerator at the United Nations' in *Proceedings of Machine Translation Summit XIV: User track*. Nice, France.
- Pouliquen, Bruno, Mazenc, Christophe, Elizalde, Cecilia and Garcia-Verdugo, Jose (2012) 'Statistical Machine Translation prototype using UN parallel documents', in *Proceedings of the 16th Annual conference of the European Association for Machine Translation*. Trento, Italy: European Association for Machine Translation.
- Pouliquen, Bruno, Mazenc, Christophe and Iorio, Aldo (2011) 'Tapta: A User-Driven Translation System for Patent Documents Based on Domain-Aware Statistical Machine Translation', in *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, 30 May 2011, Leuven, Belgium: European Association for Machine Translation.
- Prieto Ramos, Fernando (2021) 'Translating Legal Terminology and Phraseology: Between Inter-Systemic Incongruity and Multilingual Harmonization', *Perspectives* 29(2): 175–183.
- Quah, Chiew Kin (2006) *Translation and Technology*, New York: Palgrave Macmillan UK.
- Ralli, Natascia (2009) 'Terminografia e Comparazione Giuridica: Metodo, Applicazioni e Problematiche Chiave', *InTRAlinea Special Issue: Specialised Translation I*.
- Ralli, Natascia and Andreatta, Norbert (2018a) 'bistro – ein Tool für mehrsprachige Rechtsterminologie', *trans-kom* 11(1): 7–44.

- Ralli, Natascia and Andreatta, Norbert (2018b) ‘Bistro – Ein Tool Für Mehrsprachige Rechtsterminologie’, *Trans-Kom. Zeitschrift Für Translationswissenschaft Und Fachkommunikation* 11(1): 7–44.
- Ralli, Natascia and Stanizzi, Isabella (2018) ‘Il linguaggio giuridico tedesco in Alto Adige. Evoluzione delle politiche terminologiche’, *AIDAinformazioni: Rivista di Scienze dell’Informazione* 36(special issue): 169–189.
- Rei, Ricardo, Stewart, Craig, Farinha, Ana C. and Lavie, Alon (2020) ‘COMET: A Neural Framework for MT Evaluation’, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, November 2020, Online: Association for Computational Linguistics, 2685–2702.
- Reinke, Uwe (2018) ‘State of the Art in Translation Memory Technology’, in G. Rehm, F. Sasaki, D. Stein, and A. Witt (eds). *Language Technologies for a Multilingual Europe: TC3 III*, Berlin: Language Science Press, 55–84.
- Riezler, Stefan and Maxwell, John T. (2005) ‘On Some Pitfalls in Automatic Evaluation and Significance Testing for MT’, in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, June 2005, Ann Arbor, Michigan: Association for Computational Linguistics, 57–64.
- Rikters, Matiss (2018) ‘Impact of Corpora Quality on Neural Machine Translation’, in K. Muischnek and K. Müürisepp (eds). *Human Language Technologies - the Baltic Perspective - Proceedings of the Eighth International Conference Baltic HLT 2018, Tartu, Estonia, 27-29 September 2018*, 2018, IOS Press, 126–133.
- Sågvall Hein, Anna and Ekholm, Torbjörn (2020) *Language data and Machine Translation at Swedish Migration Agency (SMA)*,.
- Sandrini, Peter (1996) *Terminologiearbeit Im Recht: Deskriptiver Begriffsorientierter Ansatz Vom Standpunkt Des Übersetzers*, IITF-series ; 8., Wien: TermNet - International Network for Terminology.

- Sandrini, Peter (1998) ‘Italienisches Recht in Deutscher Sprache – Terminologische Überlegungen’, in P. Cordin, M. Iliescu, and H. Siller-Runggaldier (eds). *Italienisch Und Deutsch Im Kontakt Und Vergleich: Akten Des 7. Treffens Der Italienischen Und Österreichischen Linguisten*, 1998, Trento: Università – Dipartimento di Scienze filologiche e storiche, 399–417.
- Sandrini, Peter (2019) *Translationspolitik Für Regional- Oder Minderheitensprachen Unter Besonderer Berücksichtigung Einer Strategie Der Offenheit*, Berlin: Frank & Timme.
- Scansani, Randy (2020) *Machine translation for institutional academic texts: Output quality, terminology translation and post-editor trust*. Dottorato di ricerca in Traduzione, interpretazione e interculturalità, 32 Ciclo. Bologna: Alma Mater Studiorum Università di Bologna.
- Scansani, Randy, Bentivogli, Luisa, Bernardini, Silvia and Ferraresi, Adriano (2019) ‘MAGMATic: A Multi-Domain Academic Gold Standard with Manual Annotation of Terminology for Machine Translation Evaluation’, in *Proceedings of Machine Translation Summit XVII: Research Track*, August 2019, Dublin, Ireland: European Association for Machine Translation, 78–86.
- Scansani, Randy and Dugast, Loïc (2021) ‘Glossary Functionality in Commercial MT: Does It Help? Identifying Best Practices for an LSP’, in *Proceedings of the 18th Biennial Machine Translation Summit*, 2021, Association for Machine Translation in the Americas.
- Scansani, Randy, Federico, Marcello and Bentivogli, Luisa (2017) ‘Assessing the Use of Terminology in Phrase-Based Statistical Machine Translation for Academic Course Catalogues Translation’, in R. Basili, M. Nissim, and G. Satta (eds). *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-It 2017*, Accademia University Press, 298–303.
- Schmid, Helmut (1994) ‘Probabilistic Part-of-Speech Tagging Using Decision Trees’, in *Proceedings of International Conference on New Methods in Language Processing*, 1994, Manchester.

- Schweizerische Eidgenossenschaft (2019) *Bericht DeepL-Test: Auswertung Der Testergebnisse Und Empfehlungen Der Arbeitsgruppe 'Maschinelle Übersetzung'*. z. H. der KOSD, Bern.
- Seguin, Lucia (2021) *Machine Translation at the Government of Canada: Reaching for the Future*, MT Summit 2021.
- Sennrich, Rico, Haddow, Barry and Birch, Alexandra (2016a) 'Improving Neural Machine Translation Models with Monolingual Data', *ArXiv:1511.06709 [Cs]*.
- Sennrich, Rico, Haddow, Barry and Birch, Alexandra (2016b) 'Neural Machine Translation of Rare Words with Subword Units', *ArXiv:1508.07909 [Cs]*.
- Skadiņa, Inguna, Vasiljevs, Andrejs, Skadiņš, Raivis, Gaizauskas, Robert and Tufiş, Dan (2010) 'Analysis and Evaluation of Comparable Corpora for under Resourced Areas of Machine Translation', in *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010*, 2010, Malta: Unpublished, 6–14.
- Skadins, Raivis et al. (2020) 'Language Technology Platform for Public Administration', in U. Andrius, V. Jurgita, K. Jolantai, and K. Danguole (eds). *Human Language Technologies - The Baltic Perspective - Proceedings of the Ninth International Conference Baltic HLT 2020, Kaunas, Lithuania, September 22-23, 2020*, 2020, IOS Press, 182–190.
- Snover, Matthew, Dorr, Bonnie, Schwartz, Rich, Micciulla, Linnea and 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*, 8 August 2006, Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, 223–231.
- Soffritti, Marcello (2002) 'Die Doppelte Fachsprachlichkeit in Aktuellen Normsetzenden Texten', in M. Gotti, D. Heller, and M. Dossena (eds). *Conflict and Negotiation in Specialized Texts*, Berlin: Peter Lang.

- Somers, Harold (2011) ‘Machine Translation: History, Development, and Limitations’, in K. Malmkjær and K. Windle (eds). *The Oxford Handbook of Translation Studies*, Oxford: Oxford University Press.
- Song, Xingyi, Specia, Lucia and Cohn, Trevor (2014) ‘Data Selection for Discriminative Training in Statistical Machine Translation’, in *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, 16 June 2014, Dubrovnik, Croatia: European Association for Machine Translation, 45–52.
- Specia, Lucia et al. (2017) ‘Translation Quality and Productivity: A Study on Rich Morphology Languages’, in *Of MT Summit XVI, the 16th Machine Translation Summit*, 2017, Nagoya, Japan, 55–71.
- Stahlberg, Felix (2020) ‘Neural Machine Translation: A Review and Survey’, *ArXiv:1912.02047 [Cs]*.
- Stocker, Martha (2006) *La Storia Della Nostra Terra. Il Sudtirolo Dal 1914 al 1992 – Cenni Storici*, Bolzano: Athesia.
- Sutskever, Ilya, Vinyals, Oriol and Le, Quoc V. (2014) ‘Sequence to Sequence Learning with Neural Networks’, in *Advances in Neural Information Processing Systems*, 2014, Curran Associates, Inc.
- Tan, Zhixing et al. (2020) ‘Neural Machine Translation: A Review of Methods, Resources, and Tools’, *AI Open* 1: 5–21.
- Thüne, Eva-Maria, Elter, Irmgard and Leonardi, Simona (2011) *Le Lingue Tedesche: Per Una Descrizione Sociolinguistica*, Lecce: B.A. Graphis.
- Tiedemann, Jörg (2011) *Bitext Alignment*, San Rafael: Morgan & Claypool Publishers.
- Toral, Antonio and Sánchez-Cartagena, Víctor M. (2017) ‘A Multifaceted Evaluation of Neural versus Phrase-Based Machine Translation for 9 Language Directions’, *ArXiv:1701.02901 [Cs]*.

- Varga, Daniel et al. (2005) ‘Parallel Corpora for Medium Density Languages’, in *Proceedings of RANLP 2005*, 2005, 590–596.
- Vasconcellos, Muriel and Leon, Marjorie (1985) ‘Spanam and Engspan: Machine Translation at the Pan American Health Organization’, *Computational Linguistics* 11(2–3): 122–136.
- Vasiljevs, Andrejs, Kalniņš, Rihards, Pinnis, Mārcis and Skadiņš, Raivis (2014) *Machine translation for e-government – the Baltic case*, in 2014, 181–193.
- Vaswani, Ashish et al. (2017a) ‘Attention Is All You Need’, in I. Guyon et al. (eds). *Advances in Neural Information Processing Systems*, 2017, Curran Associates, Inc.
- Vaswani, Ashish et al. (2017b) ‘Attention Is All You Need’, *ArXiv:1706.03762 [Cs]*.
- Vauquois, Bernard (1968) ‘A Survey of Formal Grammars and Algorithms for Recognition and Transformation in Machine Translation’, in A. J. H. Morell (ed.). *Proceedings of the IFIP Congress-68*, 1968, Edinburgh: North-Holland, 254–260.
- Videsott, Paul, Videsott, Ruth and Casalicchio, Jan (2020) *Manuale Di Linguistica Ladina*, Berlin/Boston: De Gruyter.
- Vintar, Špela (2018) ‘Terminology Translation Accuracy in Statistical versus Neural MT: An Evaluation for the English-Slovene Language Pair’, in J. Du, M. Arcan, Q. Liu, and H. Isahara (eds). *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 12, -05 2018, Paris, France: European Language Resources Association (ELRA).
- Wang, Chaojun and Sennrich, Rico (2020) ‘On Exposure Bias, Hallucination and Domain Shift in Neural Machine Translation’, *ArXiv:2005.03642 [Cs]*.
- Wang, Haifeng, Wu, Hua, He, Zhongjun, Huang, Liang and Ward Church, Kenneth (2021) ‘Progress in Machine Translation’, *Engineering*.

- Wang, Weiyue, Peter, Jan-Thorsten, Rosendahl, Hendrik and Ney, Hermann (2016) ‘CharacTer: Translation Edit Rate on Character Level’, in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, August 2016, Berlin, Germany: Association for Computational Linguistics, 505–510.
- Way, Andy (2018) ‘Quality Expectations of Machine Translation’, in J. Moorkens, S. Castilho, F. Gaspari, and S. Doherty (eds). *Translation Quality Assessment: From Principles to Practice*, Cham: Springer International Publishing, 159–178.
- Way, Andy (2020) ‘Machine Translation: Where Are We at Today?’, in E. Angelone, M. Ehrensberger-Dow, and G. Massey (eds). *The Bloomsbury Companion to Language Industry Studies*, London: Bloomsbury Academic, 311–332.
- Way, Andy et al. (2020) ‘Progress of the PRINCIPLE Project: Promoting MT for Croatian, Icelandic, Irish and Norwegian’, in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, November 2020, Lisboa, Portugal: European Association for Machine Translation, 465–466.
- Wiesmann, Eva Berta Maria (2004) *Rechtsübersetzung Und Hilfsmittel Zur Translation. Wissenschaftliche Grundlagen Und Computergestützte Umsetzung Eines Lexikographischen Konzepts*, Tübingen: Gunter Narr Verlag.
- Wiesmann, Eva Berta Maria (2019) ‘Machine Translation in the Field of Law: A Study of the Translation of Italian Legal Texts into German’, *Comparative Legilinguistics* 37(1): 117–153.
- Williams, Philip, Sennrich, Rico, Post, Matt and Koehn, Philipp (2016) ‘Syntax-Based Statistical Machine Translation’, *Synthesis Lectures on Human Language Technologies* 1–208.
- Wilson, Woodrow (1918) *Address of the President of the United States: Delivered at a Joint Session of the Two Houses of Congress, January 8, 1918*, Washington, D.C.: U.S. G.P.O.

- Woelk, Jens (2000) ‘Von “Advokat” Bis “Zentraldirektion Der Autonomien”. Die Südtiroler Rechtssprache Aus Sicht Eines “bundesdeutschen” Juristen’, in D. Veronesi (ed.). *Linguistica Giuridica Italiana e Tedesca*, Padova: Unipress, 209–222.
- Wu, Yonghui et al. (2016) ‘Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation’, *ArXiv:1609.08144 [Cs]*.
- Yates, Sarah (2006) ‘Scaling the Tower of Babel Fish: An Analysis of the Machine Translation of Legal Information’, *Law Library Journal* 98(3): 481–500.
- Yngve, Victor H. (1964) ‘Implications of Mechanical Translation Research’, in *Proceedings of the American Philosophical Society*, 1964, American Philosophical Society, 275–281.
- Zafar, Iffat, Tzanidou, Giounona, Burton, Richard, Patel, Nimesh and Araujo Leonardo (2018) *Hands-On Convolutional Neural Networks with Tensorflow: Solve Computer Vision Problems with Modelling in Tensorflow and Python*, Birmingham: Packt.
- Zanettin, Federico (2002) ‘Corpora in Translation Practice’, in *Proceedings of the LREC Workshop*, 2002, 10–14.
- Zanettin, Federico (2014) ‘Corpora in Translation’, in J. House (ed.). *Translation: A Multidisciplinary Approach*, London: Palgrave Macmillan, 178–199.
- Zanon, Heinz (2001) ‘Spurensuche 1999: Die Deutsche Sprache Bei Gericht in Südtirol’, in *Die Deutsche Sprache in Südtirol. Einheitssprache Und Regionale Vielfalt*, Wien/Bozen: Folio Verlag, 166–186.
- Zanon, Heinz (2008) ‘Zur Problematik Der Entwicklung Einer Deutschen Rechtssprache Für Südtirol’, in E. Chiochetti and L. Voltmer (eds). *Normazione, Armonizzazione e Pianificazione Linguistica*, 2008, Bolzano: Eurac.

- Zarechnak, Michael (1979) ‘The History of Machine Translation’, in *Machine Translation*, Berlin/Boston: De Gruyter, 1–88.
- Zhang, Jiajun and Zong, Chengqing (2020) ‘Neural Machine Translation: Challenges, Progress and Future’, *ArXiv* abs/2004.05809.
- Zhang, Tianyi, Kishore, Varsha, Wu, Felix, Weinberger, Kilian Q. and Artzi, Yoav (2020) ‘BERTScore: Evaluating Text Generation with BERT’, *ArXiv:1904.09675 [Cs]*.
- Zoph, Barret, Yuret, Deniz, May, Jonathan and Knight, Kevin (2016) ‘Transfer Learning for Low-Resource Neural Machine Translation’, *ArXiv:1604.02201 [Cs]*.

APPENDIX A - CORPUS CLEANING AND FILTERING

1. Sentence-level cleaning

1.1 Noise at the beginning of segments

The main noise type occurring at the beginning of segments in the LEXB raw corpus are list markers, which come in the form of bullet points, dashes and several kinds of alphanumerical markers. The main kinds of noise are listed in Table 19.

type of noise	examples
bullet points	• •
dashes	- - —
asterisks	*
alphabetical markers and roman numeral markers	a. a) (a) A) (A)
numerical markers	1. 1) (1) 1.1 1.1)
other alphanumerical markers	a1) a1. 1/bis (1/bis) A1) I.A. A.1. A1

Table 19: Noise at the beginning of sentences in the LEXB corpus.

Another frequently occurring noise type at the beginning of segments is the “Art. #” string in article titles, which come in a fixed format (“Art.” + number of the article + actual title of the article between round brackets). Such segments are trimmed from both “Art. #” and round brackets, as can be seen in the examples in Table 20.

raw segment	cleaned segment
<u>Art. 1</u> (Riconoscimento di legittimità dei debiti fuori bilancio derivanti dall’acquisizione di beni e servizi)	Riconoscimento di legittimità dei debiti fuori bilancio derivanti dall’acquisizione di beni e servizi
<u>Art. 118</u> (Abweichungen für künstlerisch und geschichtlich wertvolle Bauwerke)	Abweichungen für künstlerisch und geschichtlich wertvolle Bauwerke

Table 20: Examples of raw and cleaned segments (1).

Further minor corrections of noise at the beginning of segments include noise at the beginning of text titles and opening quotation marks (if no other quotes are found in the segment). Examples are shown in Table 4:

raw segment	cleaned segment
w”) Contratto collettivo intercompartimentale 10 dicembre 2020	Contratto collettivo intercompartimentale 10 dicembre 2020
“Art. 51/ter (Reclutamento di personale da parte dell’Azienda Sanitaria dell’Alto Adige)	Reclutamento di personale da parte dell’Azienda Sanitaria dell’Alto Adige

Table 21: Examples of raw and cleaned segments (2).

1.2 Noise at the end of segments

Noise cleaned up from the end of segments includes superscripts (occurring in the form of “number + closed round bracket”), closing quotation marks (if no other quotes are found in the segment) and list markers (which are probably a result of mis-segmentation and originally referred to the following sentence). Examples are shown in Table 22.

raw segment	cleaned segment
Due membri devono appartenere al gruppo di lingua italiana e due al gruppo di lingua tedesca.46)	Due membri devono appartenere al gruppo di lingua italiana e due al gruppo di lingua tedesca.
Al riguardo si avvale anche della collaborazione di organismi nazionali e internazionali.”	Al riguardo si avvale anche della collaborazione di organismi nazionali e internazionali.
Materiale informativo sul campeggio in italiano e tedesco, nonché in una lingua straniera (3)	Materiale informativo sul campeggio in italiano e tedesco, nonché in una lingua straniera

Table 22: Examples of raw and cleaned segments (3).

1.3 Noise at the beginning and the end of segments

Noise removed from both the beginning and the end of segments includes quotation marks, square and round brackets, leading and trailing whitespaces. Brackets and quotes have been removed if both occurring at the beginning and the end of segments and if occurring at the beginning or the end of segments only.

1.4 De-hyphenation

As emerged from the preliminary manual evaluation, several sentences in the parallel corpus contained erroneously hyphenated words (e.g., “indica-zioni”), which was probably a consequence of the conversion from PDF files to plain text. A simple vocabulary-based approach was used to identify these noisy hyphenated words and correct them. The approach consists in:

2. Generating a vocabulary (frequency list) for both the Italian and German sub-corpora.
3. For any word in the corpus containing a hyphen, checking its overall frequency both as a hyphenated word and as a non-hyphenated word (if any).
4. If the non-hyphenated word is far more frequent, correcting the hyphenated word to its non-hyphenated correct version. More specifically, the approach is precision-oriented and corrects an allegedly mis-hyphenated word only if its non-hyphenated counterpart occurs at least 10x more frequently and their overall frequency in the corpus is higher than 40.

This approach proved effective and allowed to automatically correct a total number of 989 instances of mistakenly hyphenated words in the LEXB corpus.

1.5 Identical or highly similar source-target

As seen in Section 2.4.1, the presence of sentence pairs with identical source and target sentences in parallel corpora for MT training can bring about a very significant loss in MT quality. Such sentence pairs are identified and filtered out at different stages of our cleaning and filtering pipeline, in order to identify them even after they underwent sentence-internal cleaning operations (see Section 3.3.3.1).

Likewise, high source-target similarity may cause the same effect, as identical source-target sentences and can harmful to the MT system. Such pairs are identified considering the source and target Levenshtein's edit distance and edit distance ratio (edit distance normalized by the average length of source and target sentence length). Sentence pairs with a source-target edit distance lower than 2 or with an edit distance ratio lower than 0.1 (following the approach by Lu et al. 2018) were filtered out.

1.6 Non-alphabetical characters

It is assumed that sentence pairs formed mostly by punctuation or digits are not very useful for training and can therefore be eliminated (Gupta et al. 2019). Sentence pairs are therefore filtered out if the source and/or target side only contain punctuation, digits or whitespaces. Moreover, segments with a relatively high non-alphabetical to alphabetical characters ratio (>0.8) are also discarded.

1.7 Missing translation

Sentence pairs in which one of the sides is empty or contains only whitespaces are discarded. Although 1:0 or 0:1 sentence pairs with no source or no target were already discarded by LF Aligner, half-empty sentence pairs could still be found in the corpus as a consequence of previous segment-internal noise cleaning operations.

1.8 Wrong language

The wrong language filter checks whether both source and target are in the right language. As in some of the legal texts collected from the LexBrowser there are excerpts from laws or regulations in the other language, or some elements in lists are kept in the original language, it is expected that several sentences contain foreign language material (see Table 23 for examples). The filter uses the *langid* language detector (Lui and Baldwin 2012) and discards sentence pairs whose detected language does not match the expected it-de pair.

it	de
Nel testo tedesco del comma 2/bis dell'articolo 2 della legge provinciale 13 maggio 1992, n. 13, e successive modifiche, le parole “Die zertifizierte Meldung muss” sono sostituite dalle parole “Die zertifizierte Meldung der Tätigkeitsaufnahme muss” .	Im deutschen Wortlaut von Artikel 2 Absatz 2/bis des Landesgesetzes vom 13. Mai 1992, Nr. 13, in geltender Fassung, werden die Wörter „Die zertifizierte Meldung muss“ durch die Wörter „Die zertifizierte Meldung der Tätigkeitsaufnahme muss“ ersetzt.
Con deliberazione n. 457 del 18/04/2017 la Giunta Provinciale ha aggiornato i LEA nazionali sulla base di quanto disposto dal DPCM 12/01/2017 “Definizione e aggiornamento dei livelli essenziali di assistenza, di cui all'articolo 1, comma 7, del decreto legislativo 30 dicembre 1992, n. 502” .	Mit Beschluss Nr. 457 vom 18.04.2017 hat die Landesregierung die nationalen WBS gemäß der Bestimmungen im DPMR 12.01.2017 “Definizione e aggiornamento dei livelli essenziali di assistenza, di cui all'articolo 1, comma 7, del decreto legislativo 30 dicembre 1992, n. 502” aktualisiert.

Table 23: Examples of segments containing sentences in the foreign language.

2. Corpus filtering

2.1 Sentence length ratio filter

The filter removes pairs where the length ratio between source and target is higher than a certain threshold. The thresholds adopted follow the filtering approach implemented in the ModernMT open-source version.¹² The algorithm discards translation units if the source or target sentence character length exceeds the length of the translated sentence by more than 50 %. In order to prevent the filter from discarding short valid sentence pairs, an arbitrary value of 15 is added to the character count of each sentence’s length. The function for defining the sentence length ratio threshold is therefore defined as follows:¹³

$$\frac{(C_1 + 15)}{(C_2 + 15)} > 1.5$$

This allowed to identify and discard a number of bad sentence pairs resulting from actual misalignments (see Table 24).

it	de
In caso di agevolazione per l’apertura, essa deve avere luogo entro un anno dalla data di concessione del contributo, <u>fatta salva la possibilità di ottenere una proroga, per un periodo massimo di un anno, previa presentazione di motivata richiesta prima della scadenza del termine.</u>	<u>Die Frist kann jedoch vor ihrem Ablauf auf begründeten Antrag hin für höchstens ein Jahr verlängert werden.</u>
<u>Le domande per un rapporto di lavoro a tempo parziale devono essere presentate entro il mese di febbraio precedente l’inizio dell’anno formativo di riferimento, al preposto dirigente scolastica,</u> che le trasmette unitamente al proprio parere al dirigente preposto alla relativa area di formazione.	<u>Die Teilzeit-Anträge müssen innerhalb Februar vor Beginn des betreffenden Schuljahres der zuständigen Schulführungskraft vorgelegt werden.</u>

Table 24: Examples of misaligned sentences.

¹² https://github.com/modernmt/DataCollection/blob/dev/baseline/filter_hunalign_bitext.py

¹³ C_1 is the length, in terms of characters, of the longest sentence; C_2 is the character length of the shortest sentence in the pair. Whitespaces are included in the character count.

2.2 Long and short segments

Following the approach of most works related to parallel corpus filtering for MT training (see Section 2.4.1) very long and short sentences were removed from the LEXB corpus. In particular, sentence pairs with source or target segment length in terms of tokens falling outside given thresholds ($5 \geq \text{sentence length} < 80$) are discarded from the dataset.

2.3 Deduplication

The removal of perfectly duplicate sentence pairs was carried out using Heartsome TMX Editor. Near-duplicate sentence pairs were not removed during this stage, as they may be useful for MT adaptation. Still, to avoid overlapping between the adaptation set and the test set, near-duplicates were taken care of in the subsequent stage of dataset splitting (see Section 3.4.1) without actually discarding these translation units from the adaptation set.

2.4 Inconsistencies in target

Sentence pairs with the same source and different targets generate inconsistency within the training dataset and can harm MT quality. These sentence pairs are removed using Heartsome TMX Editor. When inconsistencies in target are encountered by the program, only the more recent sentence pair is kept in the corpus.

ACKNOWLEDGEMENTS

I am extremely grateful to everyone who supported me and made my journey possible.

Thanks to Prof. Adriano Ferraresi and Flavia De Camillis for their continuous support and insightful feedback at all stages of the project.

Thanks to Federico Garcea for his insights on MT adaptation and testing. Thanks to Maja Miličević Petrović for her input into the choice of the right statistical significance test.

Thanks to Daimon, Margherita and Stefan for making the last two years the best time of my life. I couldn't ask for better friends.

Thanks to Eleonora, she knows why.

Thanks to Laura and Natascia for believing in me and supporting me in my best and worst times.

Last, but not least, thanks to my family.

ABSTRACT

Following the implementation of South Tyrol's Statute of Autonomy, the public administrations of the Autonomous Province of Bozen/Bolzano are legally bound to the bilingual publication of laws and administrative acts. This results in a strong demand for translation of legal-administrative texts, usually from Italian into German, which could be satisfied, at some extent, by integrating machine translation (MT) in the institutional translation workflow. In this setting, a crucial aspect is also represented by the local South Tyrolean legal-administrative terminology, which is of central importance in institutional translation, exhibits peculiar features with respect to other German-speaking countries, and has emerged as the main issue when machine-translating Italian legal-administrative texts into South Tyrolean German.

The purpose of the present study is to adapt an MT system (ModernMT) by means of a parallel corpus of legal-administrative texts and to evaluate it both in terms of overall MT performance and in terms of legal terminology evaluation, by automatically matching and categorising the legal terms produced by the MT engine within a fine-grained taxonomy.

Results showed that the domain-adapted engine achieved a substantial and promising improvement in MT performance (+9 BLEU), yielding a relatively good score of 35 BLEU. As for legal term translation, the proposed automatic evaluation approach provided insights about terminology improvements both on a quantitative and qualitative level. The domain-adapted engine correctly translated 2746 out of 3503 legal terms of the test set (term accuracy: 78.39%), significantly outperforming the generic ModernMT system by 3.31%. The most substantial improvements were observed with regards to standardised/recommended legal terms. Despite the significant improvements in term translation accuracy, however, the adopted domain-adaptation approach did not achieve a systematic enhancement in legal terminology translation.