To cite this article: De Camillis, Flavia & Elena Chiocchetti (2024). Machine-translating legal language: error analysis on an Italian-German corpus of decrees. In Ú. Bhreathnach, N. Nissilä & A. Velicu. *Terminology Science & Research / Terminologie : Science et Recherche 27*, 1–27. Available at: https://journal-eaft-aet.net/index.php/tsr/issue/archive.

Research article

Machine-translating legal language: error analysis on an Italian-German corpus of decrees

Flavia De Camillis & Elena Chiocchetti

Eurac Research, Institute for Applied Linguistics
https://orcid.org/0000-0002-7228-5629, https://orcid.org/0000-0002-1309-7759

The paper analyzes the most frequent error categories in a bidirectional corpus of machine-translated decrees in the language combination Italian-South Tyrolean German. The aim is to assess translation issues when using a fine-tuned machine translation (MT) system to produce legal texts in an Italian province where German is an officially recognized minority language, and the local legal language differs from that used within other German-speaking legal systems. Our fine-tuned MT system struggles with features that are typical for the legal language, e.g., legal phraseology, legal terminology (especially the specific local South Tyrolean terminology), and gender-sensitive language. The latter is a requirement for local legislation. The errors identified shed light on the need to feed MT systems with terminological information, especially for low-resource language varieties such as South Tyrolean German. We consider our results key information for the training of post-editors, professional translators, and non-professional translators working in multilingual public administrations.

Keywords: machine translation evaluation, legal language, legal terminology, gender bias

Traduction automatique du langage juridique : analyse des erreurs dans un corpus de décrets italo-allemand

Flavia De Camillis & Elena Chiocchetti

Eurac Research, Institute for Applied Linguistics
https://orcid.org/0000-0002-7228-5629, https://orcid.org/0000-0002-1309-7759

Cet article analyse les catégories d'erreurs les plus fréquentes dans un corpus bidirectionnel de décrets traduits automatiquement dans la combinaison linguistique italien-allemand du Tyrol du Sud. L'objectif est d'évaluer les problèmes de traduction lors de l'utilisation d'un système optimisé de traduction automatique (TA) pour produire des textes juridiques dans une province italienne où l'allemand est une langue minoritaire officiellement reconnue, la langue juridique locale étant différente de celle utilisée dans d'autres systèmes juridiques germanophones. Notre système optimisé de traduction automatique se retrouve en difficulté face à des caractéristiques typiques du discours juridique, dont la phraséologie juridique, la terminologie juridique (en particulier la terminologie locale spécifique du Tyrol du Sud) et le langage inclusif. Ce dernier point est une exigence de la législation locale. Les erreurs identifiées mettent en lumière la nécessité d'alimenter les systèmes de traduction automatique en informations terminologiques, notamment pour les variétés de langues à faibles ressources, tel que l'allemand du Tyrol du Sud. Nous considérons nos résultats comme des acquis essentiels pour la formation des post-rédacteurs, des traducteurs professionnels ainsi que des traducteurs non professionnels travaillant dans des administrations publiques multilingues.

Mots-clés : évaluation de la traduction automatique, langage juridique, terminologie juridique, biais de genre

1 Introduction

Legal language poses notable translation challenges to humans and machines (Killman, 2023, p. 486; Mattila, 2018, p. 118). Concerning machine translation (MT), long and convoluted sentences, specialized terminology (that includes redefined general language words), and the culture-boundness of each legal system (Killman, 2023, p. 486; Kit & Wong, 2008; Mattila, 2018, pp. 122–128; Šarčević, 1997, p. 13) have made legal language generally unsuitable for MT (Sánchez-Gijón & Kenny, 2022, pp. 85–86) for a long time.

Translating legal texts is a high-stakes activity as mistakes can have serious consequences (Mattila, 2018, p. 118), such as legal disputes and infringement of basic rights (e.g., bad translation in criminal court cases). Nevertheless, MT is increasingly used to translate legal language (Mattila, 2018, p. 118; Pontrandolfo & Quinci, 2023, p. 176), including by legal translators and legal professionals (Giampieri, 2023, pp. 121–136). This holds true also for South Tyrol, Italy (De Camillis, 2021, p. 233), where German is an officially recognized minority language.

South Tyrolean German is a standard variety of German that differs from the standard varieties used in Austria, Germany, Switzerland and other German-speaking countries. Next to other minor diverging features, the main peculiarities relate to lexis in the domain of food as well as law and public administration (Ammon et al., 2016, p. LX). The latter is due to the system-boundness of legal terminology (Cao, 2007, pp. 23–24; de Groot, 1999, pp. 12–17; Šarčević, 1997, p. 13).

Citizens speaking a minority language or minority variety of a language have equal rights to access legal information and should not be discriminated against. However, speakers of these languages are not equally supported by language technologies as speakers of major languages (Rehm & Way, 2023, p. 270). In the domain of MT, there is a known quality gap when dealing with low-resource languages (Goyle et al., 2023; Ranathunga et al., 2023; Wang et al., 2021), which include small and minority languages but also minority varieties of bigger languages.

Given the growing use of MT, MT systems should be tested on minority languages and on minor varieties of big languages against their performance, particularly in high-stakes and societally relevant domains like law and administration. Any efforts to address the existing imbalance in technological support between English and other languages (Rehm & Way, 2023, p. 2) and achieve better translation quality are likely to be more successful if the issues to be tackled are well-known. Post-editing can also be guided by information on the most common (types of) MT errors.

The widespread use of MT relates also to the sustainability of terminology work. The recent neural MT technologies are not conceived to allow a straightforward integration of existing terminological resources. The different strategies that have been tested to

enforce terminology in neural MT lead to diverging results and showed notable limitations (Bane et al., 2023; Castilho & Knowles, 2024; Yvon & Rauf, 2020). Against this background, one may wonder whether it is still useful and economically sustainable to manage, update and share terminology for translation purposes in low-resource languages. Incorporating terminology into a neural MT system is a challenge and some have concluded that it is not necessarily worth the effort (Knowles et al., 2023). However, specifically in low-resource contexts, it may be important to exploit all available forms of linguistic knowledge in MT systems, including terminology (Castilho & Knowles, 2024). In future, the possibilities of prompting Large Language Models for terminology-constrained MT could make terminology resources better exploitable again (Moslem, Haque, et al., 2023; Moslem, Romani, et al., 2023).

We also see a need for in-depth analyses of MT output to avoid environmentally unsustainable MT training efforts and better focus on (other) strategies to face existing challenges. Our paper provides detailed information for the South Tyrolean context by annotating errors in a corpus of machine-translated local decrees in the Italian-German language pair.

In the following sections, we review the relevant literature (Section 2) and introduce the methodology applied to answer our research question (Section 3). The core of our paper is the presentation of the main observations (Section 4). Finally, we discuss the results (Section 5) and outline our conclusions (Section 6).

2 Literature review

Translating legal language is one of the milestones MT research and industry have not entirely reached yet, mainly due to its complexity. At the beginning of this century, MT systems still performed poorly on legal texts. Babel Fish made severe errors when translating the German and Mexican Civil Codes (Yates, 2006). Other systems (e.g., Systran) performed better in some major language combinations but did not offer fit-for-publication quality (Kit & Wong, 2008). In the legal domain, MT systems were considered adequate for gisting under certain circumstances but the aid of human translators remained essential for quality and accuracy (Mulé & Johnson, 2010).

Later research reported more promising results. Farzindar and Lapalme (2009) exploited an MT system to translate Canadian court judgments between English and French. They collected Federal Court judgments and a corpus of judgments from other institutions to train the system. After machine-translating newly published judgments and assessing the quality of the output with three human evaluators, they post-edited the outputs and measured the Post-Editing Distance, that is, the distance between raw output and post-edited version in terms of tokens and operations. Results were surprisingly good for both language directions (25% EN-FR and 23% FR-EN). Positive results were reported also by Killman (2014), who tested the translation of EU texts from Spanish into English with

Google Translate[1] (GT). He found poor syntactic output but fair terminology accuracy. 64% of the terms were translated correctly, even though the system struggled with semi-technical terms and functional words.

The studies mentioned so far all used statistical MT. Therefore, further research became necessary when neural MT became state-of-the-art and started to replace older systems in 2016. Neural MT systems achieve better fluency than previous systems (Kenny, 2022, p. 43; Killman, 2023, p. 493) but, for example, terminology constraints are more difficult to implement. Terminological accuracy can be particularly relevant when dealing with languages that are spoken in different legal systems and/or have different standard varieties (e.g., French, German, Spanish). Today, some MT systems offer the possibility of selecting a specific target language variety. DeepL[2] and GT allow users to choose between two varieties of English (British and American) and of Portuguese (Brazilian and Portuguese), while this is not yet possible for varieties of the German language.

With an eye to the different German varieties, Heiss and Soffritti (2018) assessed the performance of DeepL and GT on extracts from normative and informative texts translated from Italian to German for use in South Tyrol. They found that the specific legal and administrative terminology used in South Tyrol is one of the main issues when using MT in this region. Wiesmann (2019) tested the performance of DeepL and MateCat [3] – a web-based computer-assisted translation tool that integrates the ModernMT[4] system. Her translation tests from Italian into German included texts from the legislative area, legal practice, and legal theory. One text was aimed at a South Tyrolean audience. She found that DeepL outperformed ModernMT overall, but both systems proved generally unfit as to comprehensibility and correspondence between source and target text. Working with the Swiss variety of German, Martínez Domínguez et al. (2020) customized a system for the Swiss legal and financial language using EU parallel texts, public Swiss texts from the legal domain and specific translation memories. In this way, they achieved a higher BLEU score (12.3) compared to GT. The European institutions have also developed two systems—MT@EC first and eTranslation[5] later—that are customized for translation within the supranational EU legal system (Foti, 2022).

Other researchers presented encouraging results. Ive et al. (2020) used a web-crawled corpus of EU legal texts (e.g., disputes, procurements) and post-edited MT outputs. Then, they assessed the quality of raw MT outputs against the post-edited texts and against the reference translations, both automatically and manually. The HBLEU score of the raw machine translation compared against the post-edited version was 83 points, the BLEU score of the raw machine translation against the reference translation 32 points. A remarkable difference confirmed by human assessment, according to which

[1] https://translate.google.com/ (September 2024)
[2] https://www.deepl.com/translator (September 2024)
[3] https://www.matecat.com/ (September 2024)
[4] https://www.modernmt.com/translate/ (September 2024)
[5] https://commission.europa.eu/resources-partners/etranslation_en (September 2024)

the difference in phrasing between post-edited and reference translations did not make the latter necessarily better. Haque et al. (2020) compared legal terminology translation errors in statistical and neural MT. They found that neural MT made less terminology errors than statistical MT in the English-Hindi language pair. Error rates were lower for neural MT for both English to Hindi (8.3% vs 9.9%) and Hindi to English (11.5% vs 12.9%). This contradicts findings by Ait ElFqih and Monti (2023) who also focused on legal terminology and a low-resource language. Their comparative analysis and error evaluation on legal terminology translation between statistical and neural MT from Arabic to English and French showed that neural MT was more error-prone than statistical MT in both language pairs. They concluded that MT should be used with caution for legal content.

Quinci and Pontrandolfo (2023) studied how legal genres and levels of specialization affect the performance of MT. They worked with three legal genres (power of attorney, memorandum opinion, and share purchase agreement), two languages in combination with English (Spanish and Italian) and two legal families (common and civil law). The overall performance was not considered fully acceptable. Most translation errors concerned legal terminology and legal phraseology, in particular system-bound terms and phrases. The authors concluded that the overall acceptability of MT output depends more on the language pair than on the legal genre, probably due to the amount of available training material.

Contarino and De Camillis (2023) carried out the first proper fine-tuning for South Tyrolean legal translation exploiting ModernMT. This first adaptation process did not prove particularly promising (achieving a BLEU score of 34 points), though it showed some improvement compared to the baseline outputs of ModernMT, DeepL, and GT. This was probably due to the limited amount of South Tyrolean parallel data. The existing quality gap when using MT for low-resource languages (Goyle et al., 2023; Ranathunga et al., 2023; Wang et al., 2021) strongly relates to neural MT needing large parallel corpora to achieve good translation quality (Edman et al., 2020; Fadaee et al., 2017; Haddow et al., 2022). If these are not available, insufficient or low-quality, it is necessary, for example, to resort to monolingual data or data in other languages (Ranathunga et al., 2023) with varying results. This is why Oliver et al. (2024) tried to make up for the scarcity of South Tyrolean parallel data by combining existing data with EU parallel corpora, thus obtaining an MT system that clearly outperformed the baseline system and two commercial systems (GT and DeepL) in both translation directions.

This overview shows how MT improved considerably for the legal language over the past 20 years, especially thanks to the neural technologies and the customization process. However, we see a research gap in assessing the types of errors made by MT systems when translating legal texts. Specifically concerning terminology, research has shown that MT quality can vary notably based on the specific language pair, translation direction and MT system (cf. e.g., Haque et al., 2019, pp. 4–5 for an overview). To partly fill this gap, we answer the following research question with respect to the language pair

and low-resource language variety of our interest: What are the main translation errors when machine-translating normative texts for South Tyrol?

3 Methodology

To answer our research question, we created a bidirectional translation corpus of legal texts from South Tyrol. It was compiled in late 2021 by downloading 26 decrees in Italian and German (52 texts in total) from the textual database of local legislation LexBrowser[6]. Their average length is 1,400 tokens and the overall number of tokens is 72,000. Then, we translated 26 decrees from Italian into German and the same 26 in the opposite direction using a baseline and a fine-tuned version of ModernMT embedded in RWS Trados Studio. For the fine-tuning, we used the LEXB corpus (Contarino, 2021) and a corpus of official South Tyrolean German translations of Italian laws, for a total amount of 203,000 bilingual segments. The 26 decrees are not included in the LEXB corpus, as they have been published later. ModernMT performed surprisingly well after the fine-tuning, achieving a BLEU score of 71 points on average, calculated on both language directions. The score is so high because a notable amount of segments perfectly matched the training set due to the high repetitiveness of legal texts. Once we had excluded perfect matches from the assessment, the BLEU score still indicated a good performance (51 points on average).

The output of the fine-tuned ModernMT system was then retrieved and bundled in a bidirectional translation corpus named MT@BZ (cf. De Camillis et al., 2023). The corpus contains 26 decrees in German paired with their machine translation in Italian and 26 decrees in Italian paired with their machine translation in German. This corpus was the object of an annotation campaign. Our aim was to identify the more frequent error categories produced by the fine-tuned MT system when translating decrees in the language combination Italian-South Tyrolean German. This gave us a detailed summary of the major issues a neural MT system faces when dealing with legal language in our specific context. The annotation campaign was carried out adapting the SCATE taxonomy (Tezcan et al., 2017), which contains similar error categories to Multidimensional Quality Metrics (MQM) (Lommel et al., 2014). The campaign consisted of annotation and curation and relied on four annotators. Our gold standard corpus and annotation guidelines are available for the scientific community[7].

To assess the main translation errors, we annotated every error following our annotation scheme and performed an in-depth qualitative analysis. The annotation scheme distinguishes between Accuracy and Fluency (see Figure 1). Accuracy errors concern the transfer of meaning from source to target and are annotated on both source and target segments. Fluency errors concern the adherence to the rules and norms of the target language and are annotated solely on target segments. Both sections have

[6] http://lexbrowser.provinz.bz.it/ (September 2024)

[7] Code: https://gitlab.inf.unibz.it/commul/mt-bz/, corpus data: http://hdl.handle.net/20.500.12124/60, annotation guidelines: http://hdl.handle.net/20.500.12124/62 (September 2024)

two further sub-levels. Many categories are self-explanatory (like Addition and Omission), while some will be explained in the next paragraphs[8].

Starting from the left-hand side of Figure 1 under Mistranslation, the category Multi-Word Expressions refers to idioms, proverbs, collocations, phraseology or phrasal verbs, as well as titles of laws and decrees that should be kept together as a unit and not translated word by word. Multi-word terms (e.g., *Dekret des Landeshauptmanns, Decreto del Presidente della Provincia*) are not included in this category. The category Part-of-Speech represents an incorrect lexical category with respect to the source text (e.g., a noun in place of an adjective). Word-Sense Disambiguation and Semantically Unrelated are similar categories. In the first case, the meaning of the translated word(s) refers to a different (and wrong) meaning of the source word(s); in the second case, the meaning of the translation is not related in any way to the meaning of the source word(s). The category Partial refers to an incorrect translation due to the incomplete transmission of meaning (e.g., wrong number or gender, verb tenses and modes, etc.). Finally, Gender mistakes are incorrect translations of inclusive language (gender splitting, gender-neutral words, etc.), which the South Tyrolean administration is bound by law to use in their documents.

The category Mechanical includes Capitalization and Punctuation. It marks the wrong rendering of letter case and punctuation during translation (e.g., a comma where there was a full stop). There is an error category called Orthography that includes capitalization and punctuation errors also in the Fluency section. However, the latter refers to issues in the target language (e.g., missing capital letter at the beginning of a sentence), regardless of the source text.

Bilingual Terminology is the most relevant category for our study. For this type of error, the translation does not match the bilingual terminology requirements for South Tyrol. This means that the target term may even be a correct translation of the source term, but it is not the one used or preferred in South Tyrol.

Finally, Coherence mistakes were initially part of the scheme but have been eliminated in a later phase due to technical difficulties[9]. They are therefore grayed out in Figure 1.

[8] For a complete overview of all error definitions, see the annotation guidelines: http://hdl.handle.net/20.500.12124/62 (September 2024).

[9] In order to annotate Coherence issues, it would have been necessary to annotate mistakes on a document level. For reasons related to the annotation platform, we opted for a segment-level annotation only.

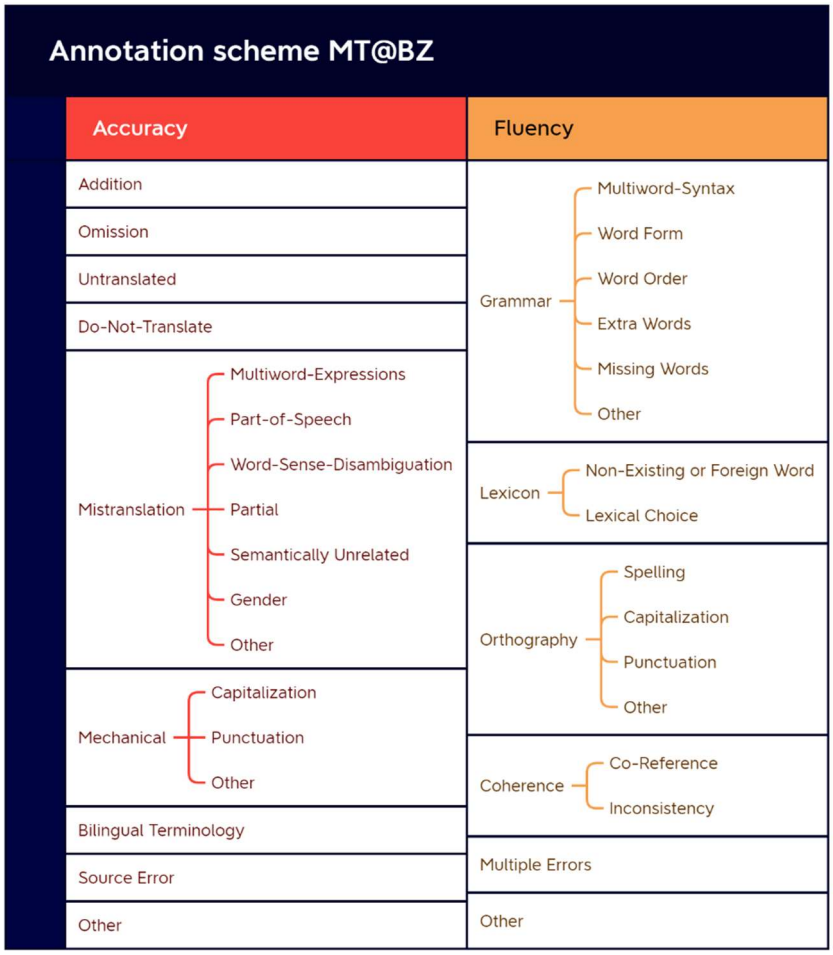


Figure 1. MT@BZ annotation scheme (source: own elaboration)

For error annotation and analysis, we used INCEpTION[10] and ANNIS[11]. ANNIS allows to perform search queries on the corpus, export the results in different formats, and run frequency analyses. We exported all annotations under each category of our scheme and further analyzed them manually in MS Excel.

For each error category, we recorded the translation direction (IT-DE or DE-IT) and additional aspects. We considered only aspects that to some extent constituted a pattern within the given category. For instance, for Addition and Omission errors, we analyzed the position of the mistake within the segment (beginning, middle, end), the length of the segment (short, medium, long), and the presence or absence of a verb in the segment, as well as its form (finite vs non-finite). For Gender errors, we only took note of the translation strategy (i.e., a gender-splitting translated with a masculine form, a gender-neutral noun translated with a masculine noun, etc.). For Multi-Word Expressions, we annotated whether the expression was a title, a set expression, or a quote.

[10] https://inception-project.github.io/ (September 2024)
[11] https://corpus-tools.org/annis/ (September 2024)

4 Observations

As reported in Figure 2, the majority of annotations concern Accuracy (2,516 errors), not Fluency (723 errors). This confirms that neural MT systems produce much more fluent translations than previous technologies. Mistranslations are the most represented sub-category within Accuracy. Bilingual Terminology, Addition, Omission, and Mechanical errors follow. Other Accuracy sub-categories are present only residually.

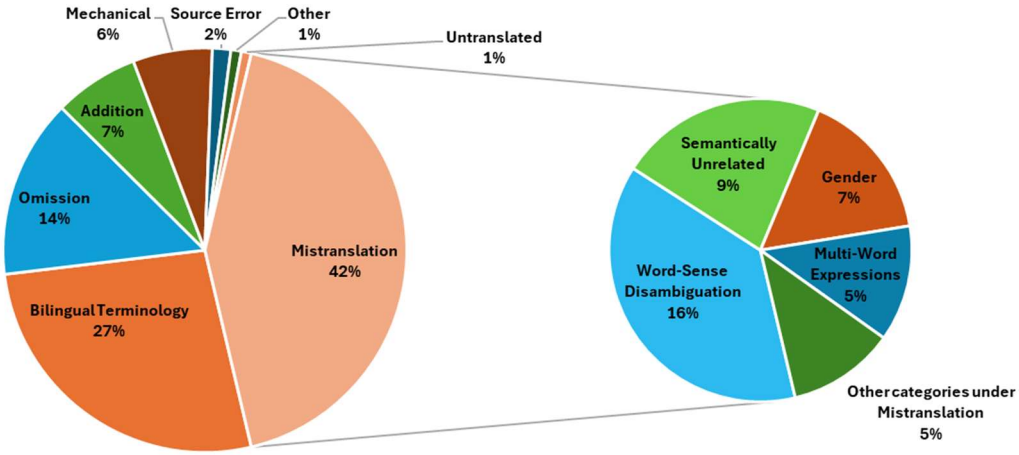


Figure 2. Accuracy errors by main category and by Mistranslation subcategory (source: own elaboration)

As for Fluency issues (see Figure 3), there are mostly Grammar, Orthography, and Lexicon errors.

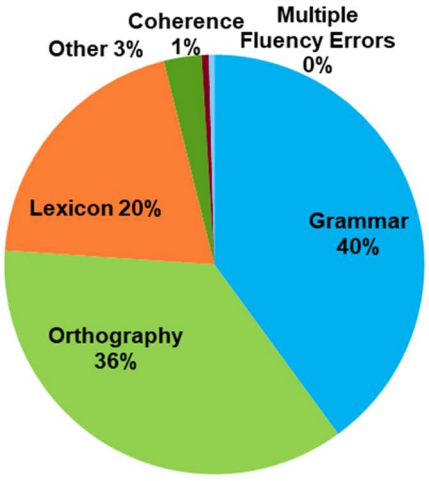


Figure 3. Fluency errors by category (source: own elaboration)

In this section, we will focus only on some categories with potentially interesting translation phenomena related to the legal language. All observations are related to our specific corpus, text type, MT system, and language directions. The numbers are small and our results are not necessarily generalizable. However, they can suggest potential paths for further analyses.

4.1 Addition

In our corpus, we annotated 171 instances of Addition. 109 were errors in the DE-IT translation direction, only 62 in the opposite direction. Often, short sentences and segments without a verb, such as titles or items of a list, had extra words at the very beginning or at the very end of the segment. In many cases, these extra words were numbers (see Example 1) or words that thematically fit the legal genre[12]. Occasionally, they are articles (Example 2). In Example 3, *considerando quanto segue* is a recurrent phrase in Italian decrees, but it is never placed after the number of the paragraph.

Example	*Source segment*	*MT output*
1	Anwendungsbereich	**198.** Ambito di applicazione.
2	Artikel 10	**L'**articolo 10;
3	Artikel 4	Articolo 4 **considerando quanto segue:**

In longer and more complex sentences, additions do not stand out for their inappropriateness; they usually fit well in the sentence. Sometimes, a noun or a phrase that is already present in the sentence is added in an identical or similar form, creating a repetition or hallucination (see Example 4, where *eines Experten oder Experten oder eines Vertreters einer Berufsorganisation* is repeated).

Example	*Source segment*	*MT output*
4	Nei casi in cui, a causa dello stato di emergenza, nello svolgimento dell'esame finalizzato al conseguimento del diploma non viene svolto uno stage è possibile derogare dalla nomina a membro della commissione d'esame, ai sensi dell'articolo 16, comma 1, lettera d), della deliberazione della Giunta provinciale 9 ottobre 2018, n. 1027, di un'esperta o un esperto ovvero di un rappresentante di un'organizzazione professionale.	In Fällen, in denen aufgrund des Notstandes bei der Durchführung der Diplomprüfung kein Praktikum stattfindet, kann von der Ernennung eines Experten oder Experten oder eines Vertreters einer Berufsorganisation als Mitglied des Prüfungsausschusses gemäß Artikel 16 Absatz 1 Buchstabe d des Beschlusses des Provinzialrats vom 9. Oktober 2018, Nr. 1027, **eines Experten oder Experten oder eines Vertreters einer Berufsorganisation** abgewichen werden.

[12] All words appearing in boldface in the tables indicate the parts of the sentences that the MT system wrongly added, omitted or mistranslated.

In our dataset, additions in longer sentences are more frequently located in the middle of the segment rather than at the beginning or at the end. This suggests that this type of mistake is more likely to occur in complex sentences rather than in simple and short sentences.

4.2 Omission

In our corpus, we found 211 instances of Omission in the IT-DE direction and 150 in the DE-IT direction (361 overall). Generally, words were omitted in segments with a finite verb, usually from the middle of the sentence and in shorter or medium-length segments. Nouns or nominal groups, adjectives, parts of sentences, and verbs or verbal groups were omitted more frequently. Function words such as conjunctions, prepositions, and determiners were seldom omitted. As for error severity, we used a three-level scale—high, medium, low—according to how much the meaning of the sentence was compromised. We categorized almost half of the omissions as low severity, while the other half is split into medium and high severity.

We noticed that omissions of medium or high severity usually included nouns, verbs, or whole pieces of the sentence, while they involved adjectives, conjunctions, and numbers more rarely. The latter categories were more represented in low-severity omissions, together with special characters (e.g., %) and adverbs. There are exceptions to this pattern, as shown in Examples 5 and 6. In Example 5, the prepositional group *in caso di* represents a condition that gets lost in the translation process: In case a certain facility is used, users are entitled to a 30% reduction. However, the condition disappears in the German version and what is reduced is not the amount due but rather the degree of use of the facility.

In Example 6, the present participle *mitarbeitende* carries important information that changes the meaning of the sentence considerably. The subjects of the sentence have to pay social security contributions for family members who work with them and have no other social security coverage, not for any member of the family, as the Italian translation states.

Example	*Source segment*	*MT output*
5	a) riduzione del 30 per cento **in caso di** utilizzo di impianti consortili o gestiti in comune sotto altra forma giuridica;	a) Verringerung der Nutzung von gemeinschaftlich genutzten Einrichtungen oder Einrichtungen anderer Rechtsform um 30%;
6	Sie sind zudem verpflichtet, die Rentenbeiträge für **mitarbeitende** Familienmitglieder, die nicht anderweitig rentenversichert sind, einzuzahlen.	Essi devono inoltre versare i contributi previdenziali per tutti i familiari che sono privi di altra assicurazione.

Omissions of verbs can impact the meaning more seriously than nouns. Sometimes, the entire sense of the sentence is compromised, as shown in Example 7. From the Italian target sentence, it is not clear whether the installments must be paid or not, while the German source clearly states that the payment of installments is suspended.

Example	*Source segment*	*MT output*
7	Auf Anfrage der einzelnen Bauspardarlehensnehmer **wird** ab dem 1. Jänner 2021 eine Aussetzung der Rückzahlung der Darlehensraten für einen Zeitraum von maximal 12 Monaten bis zum 31.12.2021 **gewährt**.	Su richiesta del singolo mutuatario, Gennaio 2021, la sospensione del rimborso delle rate di finanziamento per un periodo massimo di 12 mesi fino al 31.12.2021.

Nouns or nominal groups were omitted much more often than other elements. This can be due to them being the most represented part of speech in the legal and administrative language. Many are legal terms. Chromá (2008, p. 311) calculated how many terms are contained in different legal texts, reaching percentages between 20% and 29%. However, not all omissions of nouns or nominal groups had serious consequences. We found cases where the machine simplified legalese, resulting in a useful reduction of redundancy (Example 8).

Example	*Source segment*	*MT output*
8	La richiesta scritta va inoltrata alla Ripartizione provinciale Politiche Sociali, Servizio di valutazione della non autosufficienza e deve contenere **tutte le informazioni relative alle** variazioni della situazione assistenziale.	Im Antrag an die Landesabteilung Soziales, Dienst für Pflegeeinstufung, müssen die erfolgten Änderungen in der Betreuungssituation beschrieben sein.

4.3 Bilingual terminology

Terminology is one of the primary sources of difficulties in legal translation (Cao, 2007, p. 53; Killman, 2023, p. 487; Prieto Ramos & Cerutti Benitez, 2021, p. 156). The legal terminology used in South Tyrol can differ from the terminology used at the EU level in Italian and German and from German-language terminology used at the national level in Germany, Austria, and Switzerland. In particular for German, we must consider that, in South Tyrol, the legal system of reference is the Italian one, which is expressed in an official minority language at the local level (Chiocchetti, 2021). Legal terminology reflects the legal specificities of each legal system and is not only culture-bound but strongly system-bound. With respect to other legal systems using German as an official language, South Tyrolean German terminology is characterized by: i) unique legal terms, ii) terminological variants across legal systems, and iii) homonyms that designate

different concepts across legal systems. These categories apply to most situations in which the same official language is used to express different legal systems.

Unique legal terms are (mainly German) legal terms that are unknown or scarcely used elsewhere outside South Tyrol. This group includes terms that designate bodies, institutions, and specific Italian legal concepts at the local or national level (e.g., *Anwaltschaft des Landes, Bereichsübergreifender Kollektivvertrag, Azienda sanitaria dell'Alto Adige*).

The second group consists in terms that designate legal concepts for which equivalent or at least largely comparable legal concepts exist in other legal systems using the same official language. Despite the conceptual equivalence, the designations are different in various legal systems. In other words, due to the system-boundness of legal language, there is terminological variation (Freixa, 2022) across legal systems (e.g., South Tyrolean legal texts are subdivided into *Abschnitte* rather than *Kapitel* as in Switzerland and further into *Artikel* rather than *Paragraphen* as in Germany). This phenomenon operates also across legal levels within the same language (e.g., *Amtsblatt* corresponds to *Gazzetta Ufficiale* at the Italian national level but to *Bollettino Ufficiale* at the local level).

Homonyms that designate different legal concepts are a frequent feature across legal subdomains even within the same legal system (e.g., a "bank" is an institution in financial law but also the assembly of all the judges of a court in procedural law). If we analyze the phenomenon across legal systems, we find that the same term may be used to refer to (completely) different concepts (e.g., while *Richtlinien* corresponds to *direttive* at the EU level, it corresponds to *criteri*, a different type of legal text, in most cases where the term is used in South Tyrol).

In addition, in South Tyrol, there is a set of officially validated terms that designate Italian legal concepts (Chiocchetti, 2021). Approximately 7,400 couples of designations in Italian and German have been officially validated by a Terminology Commission. Public institutions have the legal obligation to use the standardized terminology (e.g., a collective bargaining agreement is a *contratto collettivo* in Italian and a *Kollektivvertrag* in German but never a *Tarifvertrag* as it would be in Germany). In 2012, the Terminology Commission suspended their activity but a further set of about 350 designations was recommended for use in South Tyrol by the Office for Language Issues of the provincial administration.

When starting our project, we presumed that ModernMT was more likely to have been trained with freely available EU legal texts and with larger amounts of texts from the main German-speaking countries than with many South Tyrolean texts. Our annotation results seem to support this assumption since Bilingual Terminology is our second-largest error category. In our corpus, we annotated 674 errors (306 DE-IT, 368 IT-DE), slightly more from the national language Italian into the minority language German than in the other translation direction. Some of these errors (31 occurrences) would actually

have been acceptable translations in the context of another legal system (e.g., in Austria, Germany, Switzerland, or at the EU level) but not in South Tyrol (see Example 9 where *direttiva* corresponds to *Richtlinien* at the EU level but should be translated with *criteri* at the local South Tyrolean level). We also found 129 occurrences where the officially standardized terminology had been disrespected (see Example 10 where the correct and standardized term is *Beschluss* and not *Entschließung*).

Example	*Source segment*	*MT output*
9	Die Studienbeihilfen laut diesen **Richtlinien** sind mit keiner anderen Studienbeihilfe für dieselbe Ausbildung kumulierbar.	Gli assegni di studio di cui alle presenti **direttive** non sono cumulabili con altri assegni di studio per la stessa formazione.
10	Il Piano di sviluppo dei servizi per la prima infanzia di cui all' articolo 2 dell' allegato A alla **deliberazione** n. 666 del 30 luglio 2019 non deve essere presentato.	Der in Artikel 2 des Anhangs A der **Entschließung** Nr. 666 vom 30. Juli 2019 genannte Entwicklungsplan für frühkindliche Dienstleistungen ist nicht vorzulegen.

Equally problematic were obsolete translations (9 occurrences). For example, the designation used to refer to the decrees of the president of the province was officially changed in 2001 from *Decreto del Presidente della Giunta provinciale* to *Decreto del Presidente della Provincia*. Given that the system was fine-tuned with texts containing both the current and the obsolete term, this is likely to have caused the mistake.

Many Bilingual Terminology errors derived from literal translations of legal terms (see Example 11 where both *criteri* and *microstrutture* were translated literally and not with *Richtlinien* and *Kindertagesstätten* respectively).

Example	*Source segment*	*MT output*
11	I **criteri** per il finanziamento delle **microstrutture** e del servizio di assistenza domiciliare all'infanzia/Tagesmütter sono stabiliti nell'allegato A della deliberazione della Giunta provinciale n. 666 del 30 luglio 2019.	Die **Kriterien** für die Finanzierung von **Mikrostrukturen** und der häuslichen Pflege für Kinder/Tagesmütter sind in Anlage A des Beschlusses des Provinzrats Nr. 666 vom 30.

Finally, 12 Bilingual Terminology errors concerned abbreviations (e.g., *suppl. ord.* for *supplemento ordinario*), acronyms (e.g., *IMI* for *imposta municipale immobiliare*) and initialisms (e.g., *ELR* for *Entwicklungsprogramm für den ländlichen Raum*), including their occurrences in compounds (e.g., *IMI-Befreiung* used in place of *GIS-Befreiung*). Handling short forms is a known limitation of MT systems (Sánchez-Gijón & Kenny, 2022, p. 89) but they are frequent in legal texts. Short forms were either untranslated (see Example 12 where *B.U.* for *Bollettino Ufficiale* remains unchanged but should have become *A.Bl.* for *Amtsblatt*), replaced with an invented acronym or translated with the wrong full

form (see Example 13 where *D.P.P.* is the initialism for *Decreto del Presidente della Provincia*, in German *Dekret des Landeshauptmanns*, not *Präsidialerlass*).

Example	*Source segment*	*MT output*
12	Pubblicato nel supplemento 2 al **B.U.** 19 novembre 2020, n. 47.	Veröffentlicht in Beilage 2 zur **B.U.** Nr. 47 vom 19. November 2020.
13	L'art. 4, comma 3, è stato così sostituito dall'art.1, comma 1, del **D.P.P.** 4 febbraio 2021, n. 4.	Art. Absatz 3 wurde so durch Art. 4 Absatz 3 ersetzt. 1 Absatz 1 des **Präsidialerlasses** 4. Februar 2021, Nr. 4.

4.4 Mistranslation: Multi-Word Expressions

We observed 133 mistranslations of multi-word expressions (81 for IT-DE and 52 DE-IT): 2 were literal quotes from previously published decrees; 45 titles of decrees or laws; 86 instances of phraseology. This type of error is particularly interesting for legal language analyses, as recurrent phraseology and intertextual references are common in this genre (Mattila, 2018, pp. 124–125). Examples 14 and 15 show two frequent expressions of the Italian legal language: *e successive modifiche* and *convertito in legge, con modificazioni*. From a semantic point of view, the literal translations are correct. However, they are not the conventional phraseology used in South Tyrol (*in geltender Fassung* and *mit Änderungen umgewandelt durch das Gesetz*).

Example	*Source segment*	*MT output*
14	"Criteri per il riconoscimento dello stato di non autosufficienza e per l'erogazione dell'assegno di cura" di cui all'Allegato 1 della deliberazione della Giunta provinciale n. 1246 del 14 novembre 2017, **e successive modifiche**	Richtlinien zur Anerkennung der Pflegebedürftigkeit und zur Auszahlung des Pflegegeldes laut Anlage 1 zum Beschluss der Landesregierung Nr. 1246 vom 14. November 2017 **und nachfolgende Änderungen**
15	Per il conteggio del limite di 10 giorni di calendario previsto all'articolo 18, comma 13, non vengono considerate le giornate aggiuntive di aspettativa retribuita ai sensi del decreto "Cura Italia" (decreto-legge 17 marzo 2020, n. 18, **convertito in legge, con modificazioni,** dalla legge 24 aprile 2020, n. 27) o di altre misure similari.	Für die Berechnung des Limits von 10 Kalendertagen gemäß Artikel 18 Absatz 13 werden die zusätzlichen Tage bezahlten Urlaubs gemäß dem Dekret „Cura Italia" (Gesetzesdekret Nr. 18 vom 17. März 2020, **mit Änderungen, durch Gesetz** Nr. 27 vom 24. April 2020) oder anderen ähnlichen Maßnahmen nicht berücksichtigt.

A similar phenomenon occurs with intertextual references. Despite the training with a corpus of local legislation, the machine was not able to retrieve the title of a previously published piece of legislation correctly. It translated the title from scratch, as shown in Example 16.

Example	*Source segment*	*MT output*
16	Criteri **"Disciplina di autorizzazione e accreditamento dei servizi sociali e sociosanitari"** di cui all'Allegato A della deliberazione della Giunta provinciale n. 535 del 25 giugno 2019, e successive modifiche	Kriterien **„Regelung der Genehmigung und Akkreditierung sozialer und sozialer Dienste"** gemäß Anlage A des Beschlusses des Provinzrats Nr. 535 vom 25.

4.5 Mistranslation: Gender

We identified 173 errors related to gender-sensitive language (119 in the DE-IT direction, 54 in the IT-DE direction). This imbalance in the distribution suggests that the German texts in the training corpus may contain more instances of non-sexist formulations than the Italian ones. In fact, the German language has a longer tradition of using gender-sensitive language in texts. Since 2010, South Tyrolean public institutions are bound by law to avoid sexist language in legal and institutional documents (Provincial Law No. 51/2010). We assume that provincial documents drafted later are likely to contain strategies for gender-sensitive writing, such as gender-splitting (i.e., using both the masculine and feminine form), gender-neutral nouns or collective nouns, but also that most previous publications are likely to contain masculine nouns used as a generic reference to anyone (generic masculine). Since the training set included texts from the 1970s to 2020, it must be quite inconsistent as far as non-sexist language is concerned. It was therefore to be expected that the machine would struggle with gender-sensitive translations. Interestingly, the machine was not even able to handle gender-splitting properly, despite the explicit presence of a masculine and feminine noun in the source sentence.

We observed 96 errors related to gender-splitting (see Example 17) and 61 errors of gender-neutral nouns translated only with a masculine form (Example 18). 10 errors are interesting from a technical point of view because the machine translated a splitting into a "loop", that is, a repeated masculine noun (Example 19).

Example	*Source segment*	*MT output*
17	Die Auszahlung des zustehenden Beitrages wird auf Grundlage der im Antrag angeführten Erklärungen **vom Direktor/von der Direktorin** des zuständigen Landesamtes verfügt.	La liquidazione del contributo spettante è disposta **dal direttore** dell'ufficio provinciale competente sulla base di quanto dichiarato nella domanda di contributo.
18	d) Kopie der Identitätskarte **der Antrag stellenden Person**	d) copia della carta d'identità **del richiedente**
19	In presenza di gravi motivi **la direttrice o il direttore** del Comprensorio, sentito **la direttrice o il direttore** del servizio territorialmente competente dell'Azienda Sanitaria e **la direttrice o**	Bei Vorliegen schwerwiegender Gründe kann der **Direktor oder der Direktor** des Bezirks, nach Rücksprache mit **dem Direktor oder dem Direktor** des örtlich zuständigen Dienstes der Gesundheitsbehörde und **dem Direktor**

	il direttore della residenza per anziani, può provvedere in qualsiasi momento, senza preavviso, alla revoca dell'incarico.	**oder dem Direktor** des Pflegeheims, die Bestellung jederzeit widerrufen.

We also assessed whether the conjunction or symbol used within the splitting in the source segment (*/, e, o, und, oder, bzw.*) had an impact on the translation but did not find any pattern.

4.6 Mistranslation: Word-Sense Disambiguation and Semantically Unrelated

We annotated 405 instances of Word-Sense Disambiguation errors and 238 of Semantically Unrelated errors. These two categories are very similar. They are both located along the *continuum* of meaning equivalence. Let's imagine this *continuum* as a circle. In the middle of the circle, there is a correct translation equivalent for a given source word in a given context. Moving from the middle towards the border of the circle, there are other words that may translate the source but in other contexts (Word-Sense Disambiguation). They are possible translations, just not in the given context. Outside the circle, there are all words that cannot be considered translations of the source in any possible context (Semantically Unrelated). Since all Word-Sense Disambiguation and Semantically Unrelated mistakes are placed along this *continuum*, we struggled to further categorize them besides the original separation "in" and "out" of the circle. There definitely are small differences among the instances, which, however, do not show clear patterns. For this reason, the following description is anecdotal rather than systematic.

Clear-cut examples of Word-Sense Disambiguation errors are, for instance, *Absatz* translated with *paragrafo*, which is a section of a general text, rather than with *comma*, the correct term for a specific section of a South Tyrolean legal text (Example 20). Similarly, *sezione* translated with *Abschnitt* instead of *capo* is the wrong designation for a section of the Italian criminal code (Example 21).

Example	*Source segment*	*MT output*
20	e) für die Fahrkartenausgabe- und -lesegeräte laut Artikel 5 **Absatz** 1 Buchstabe c) eine Erklärung des/der Begünstigten über deren Installation.	e) per l'emissione del biglietto e per i lettori di cui all'articolo 5, **paragrafo** 1, lettera c), una dichiarazione del beneficiario o dei beneficiari relativa alla loro installazione .
21	Sie dürfen nicht verurteilt worden sein, auch nicht mit noch nicht endgültigem Urteil für eine der Straftaten laut II. Buch II. Titel I. **Abschnitt** des Strafgesetzbuches,	Non devono essere stati condannati, neppure con sentenza passata in giudicato per uno dei reati previsti dalla II. Libro II. Titolo I. **Sezione** del codice penale,

All wrong translations would be perfectly fine in other contexts. Possibly, they actually were correct in other contexts contained in the training set and landed in the output for

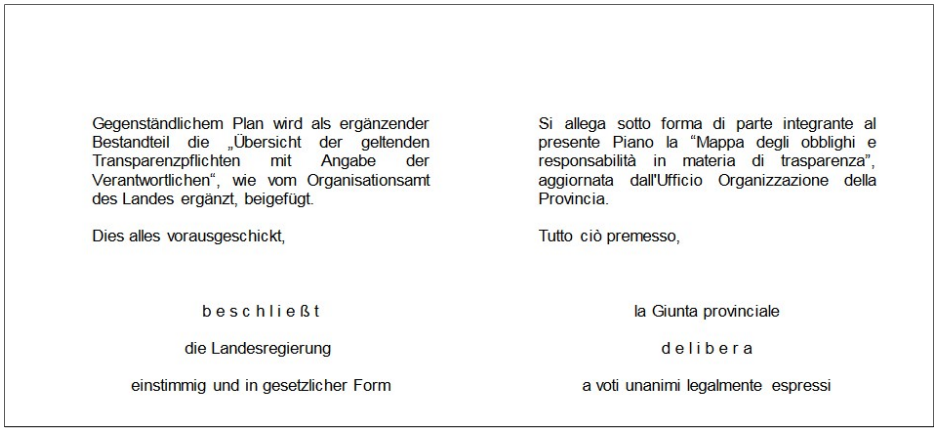
this reason. Another example of Word-Sense Disambiguation mistakes are literal translations that end up referring to another concept or to translations that are not idiomatic (Example 22). We may suppose that, in the absence of substantial results for a specific word or sequence of words, the machine stayed on the safe "literal" side, as in the case of *disciplina di missione* translated with *Missionsdisziplin* (which is a religious "mission", not a "business trip").

Example	*Source segment*	*MT output*
22	La **disciplina di missione** vigente per il personale dell'Amministrazione è stata sottoscritta in data 09.04.2008 e aggiunta come allegato 1 al Contratto collettivo intercompartimentale per il periodo 2005-2008 per la parte giuridica e per il periodo 2007-2008 per la parte economica.	Die derzeitige **Missionsdisziplin** für Verwaltungspersonal wurde am 09.04.2008 unterzeichnet und für den Zeitraum 2005-2008 für den rechtlichen Teil und für den Zeitraum 2007-2008 für den wirtschaftlichen Teil als Anhang 1 zum Kollektivvertrag zwischen den Abteilungen hinzugefügt.

Different scenarios unfold for Semantically Unrelated errors. Many are surprisingly bizarre mistakes, since the source text words were highly represented in the training corpus and often translated correctly. Nevertheless, the machine sometimes produced random translations. This is the case for *Anlage* translated with *tabella* instead of *allegato* ("table" instead of "annex", Example 23); *articolo* translated with *Kapitel* instead of *Artikel* ("chapter" instead of "article", Example 24); 9 translated with 8, and many more mistranslated numbers (Example 25). These words are likely to be found near the actual equivalent or in similar contexts: An annex might contain a table; a chapter might be split into articles; 9 comes after 8. Thus, we might speculate that this group of Semantically Unrelated errors originate in a misalignment between source and target in the training.

Example	*Source segment*	*MT output*
23	**Anlage** A	**Tabella** A
24	**Articolo** 2	**Kapitel** II
25	Art . **9**	Artikel **8**

Other Semantically Unrelated errors could derive from misalignment. This is the case for *la giunta Provinciale* translated with *dies vorausgeschickt* or *a voti unanimi legalmente espressi* translated with *die Landesregierung*. As shown in Figure 4, these sentences usually appear in an inverted order in the Italian and the German version, due to the inherent structure of each language.



Gegenständlichem Plan wird als ergänzender Bestandteil die „Übersicht der geltenden Transparenzpflichten mit Angabe der Verantwortlichen", wie vom Organisationsamt des Landes ergänzt, beigefügt.

Dies alles vorausgeschickt,

b e s c h l i e ß t

die Landesregierung

einstimmig und in gesetzlicher Form

Si allega sotto forma di parte integrante al presente Piano la "Mappa degli obblighi e responsabilità in materia di trasparenza", aggiornata dall'Ufficio Organizzazione della Provincia.

Tutto ciò premesso,

la Giunta provinciale

d e l i b e r a

a voti unanimi legalmente espressi

Figure 4. Excerpt from Provincial Council Decision No. 317/2019

Some Semantically Unrelated errors appear to be very similar to source language words. One example is the *cumulabile* translated with *kumulierte* instead of *kumulierbar* ("cumulated" instead of "cumulable", Example 26).

Example	*Source segment*	*MT output*
26	e) se un medico opera esclusivamente in più strutture di cui ai punti a) e b), il trattamento economico massimo **cumulabile** corrisponde ad Euro 2.700,00 (duemilasettecento//00) mensili; in altri casi il trattamento economico massimo **cumulabile** corrisponde ad Euro 3.000,00 (tremila//00) mensili.	e) wenn ein Arzt ausschließlich in mehreren unter den Buchstaben a) und b) genannten Einrichtungen arbeitet, entspricht die maximale **kumulierte** wirtschaftliche Behandlung 2.700,00 Euro (zweitausend siebenhundert/00) pro Monat; in anderen Fällen entspricht die maximale **kumulierte** wirtschaftliche Behandlung 3.000,00 Euro (dreitausend/00) pro Monat.

Finally, the last type of Semantically Unrelated errors are nonsensical translations. They are more frequently encountered in short segments like titles or list items. In such cases, the machine produced an entirely different and out-of-context translation for the source content (see Examples 27, 28, and 29).

Example	*Source segment*	*MT output*
27	a) vigila sulla regolare tenuta e conservazione della documentazione sanitaria;	(2) Die Kommission
28	Kontrollen	Liquidazione del contributo 7.
29	Articolo 5	Soweit in den Einzelrichtlinien Grenzwerte aufgeführt sind, besteht keine Abweichung von dem genehmigten Typ, wenn diese Grenzwerte eingehalten werden.

5 Discussion

Before delving into the discussion, we would like to acknowledge that our interpretation of the results is strongly anchored in (and limited by) our translation background and does not aim at giving detailed technical explanations.

First, we observed that, in our dataset, words or parts of the sentence are more frequently added or omitted at the beginning or at the end of short or incomplete segments (e.g., titles or items of a list). Quite differently, in longer segments extra words tend to be added or omitted more frequently in the middle. As for additions, in shorter sentences, the extra words seem quite random, though they fit the legal context. In longer sentences, additions are often iterations, as the machine repeats a part of the sentence. For longer sentences, repetitions may be due to the different syntactic order in Italian and German and to sentence length. These two factors may have led the machine to lose track of the inner structure of the sentence, while mostly preserving the overall meaning. As for omissions, half were of low severity. Omitted nominal and verbal groups represented the most severe omissions. However, also adjectives and functional words like prepositions can carry an important message and omitting them can considerably alter the overall meaning.

Second, Bilingual Terminology errors are particularly relevant for the legal language of South Tyrol. Terminology makes up a notable part of legal texts and is one of the main sources of difficulties in legal translation. Not all terminology mistakes impair the comprehension of the text but they all affect its legal validity. While an incorrect term is acceptable when MT is used for gisting, it is unacceptable to use invented literal translations, obsolete terms, or even terms from other legal systems within a law or any other legal text aimed for publication. Mistakes in legal terminology can have far-reaching consequences. As we have experienced with our experiments on South Tyrolean German, low-resource varieties of major languages also face specific challenges. In our case, despite the good results in terms of automatic scores, it was not sufficient to fine-tune an MT system with most of the available bilingual legislation for South Tyrol to eliminate all terminology mistakes, quite likely because the amount of text was not sufficient and the range of topics treated was too wide for efficient learning. Consistency within the training corpus was also an issue, as the occurrences of obsolete terms seem to suggest.

Multi-Word Expressions is a particularly interesting error category, as it includes legal phraseology and intertextual references, two very recurrent elements in normative texts. The machine could never retrieve the title of a previously published piece of law from its training material. For phraseology, we observed the same phenomena as for legal terminology. This means that the errors derive from the inability of the machine to learn and apply the local linguistic conventions, in particular in the local variety of German. This may be due to an excessively diluted presence of these expressions in the training material.

As for Gender errors, we can confirm the tendency to "male default" translation with the ensuing under-representation of other genders (Savoldi et al., 2021). Like Triboulet and Bouillon (2023), we found this tendency especially for MOFCs (Male Occupation in Female Context, such as "This mechanic is very serious when it comes to her work."), while we had a systematic mistranslation of feminine instances both for stereotypically male occupations (e.g., director and expert) and for non-stereotyped roles (e.g., student and employee). Previous research has illustrated gender bias in translation between English and languages with grammatical gender (Hovy et al., 2020; Stanovsky et al., 2019), while studies on translation between languages other than English are still limited (Costa-jussà, 2019, p. 497; Monti, 2020, p. 463). Like other researchers (Cabrera & Niehues, 2023; Marzi, 2021), we found gender errors when translating between two languages that are equally marked in terms of grammatical gender (Italian and German). This leads us to discard the assumption that gender bias mainly emerges when it comes to languages with an asymmetrical treatment of gender. Nonetheless, we cannot completely dismiss the possibility that ModernMT translated between Italian and German through English. The presence of a pivot language that is less marked in terms of grammatical gender could explain one of the phenomena we observed in our corpus, that is, the repetition of the masculine term.

Finally, we saw how Word-Sense Disambiguation and Semantically Unrelated errors are two adjacent categories, which made it particularly hard to differentiate between them. Nonetheless, we detected several forms of semantic errors in our dataset: from possible correct meanings of the source word but in the wrong context, to literal translations for uncommon source words, to complete nonsensical translations, including the incorrect rendering of numbers. On the one hand, especially for mistakes related to a wrong meaning of the word as well as nonsensical translations, we can assume that the context was not clear or long enough for the machine to disambiguate correctly. On the other hand, literal solutions point to an insufficient amount of occurrences in the corpus, hence—once again—to an insufficient corpus size.

6 Conclusions

In this paper, we identified the main translation errors when machine-translating legal texts for South Tyrol. Our in-depth qualitative analyses confirmed that legal terminology remains a key factor for the correct translation for South Tyrol and an open issue for MT. We were also able to identify further, more subtle issues not focused on in previous studies. For instance, we have seen how serious additions and omissions may be, depending on what is added or omitted in which position in the sentence, and how difficult it can be to spot them. In addition, gender mistakes turned out to be crucial for legal translation in South Tyrol, due to the obligation for public institutions to use non-sexist language. Also, legal phraseology and titles of laws, mostly translated literally rather than replaced with their correct version in the target language, represent a major issue due to the high formality and intertextuality of legal texts. Finally, context-related information seems decisive not only for properly machine-translating legal terms, but also terms that are rare or polysemous.

We are aware that our results may strongly depend on the MT system we used and the fine-tuning we performed. However, the fine-tuning with South Tyrolean legal texts yielded good results. A substantial amount of repetitive segments was automatically retrieved without the need for re-translation and the remaining segments were of good quality.

The errors in our dataset confirm two well-known issues. First, there is an evident need to feed terminological information into MT systems. This does not only relate to pairs of terms. Context-related information would help disambiguate synonyms, homonyms, and polysemic terms. This is particularly relevant for a minority language variety such as South Tyrolean German. Second, translation errors produced by neural MT systems are often subtle and easily overlooked. This may not be problematic for trained post-editors, but it certainly is for non-professional translators in the South Tyrolean administration. Therefore, we consider our results key information for the training of both post-editors and non-professional translators.

As concerns the sustainability of future terminology work, we think that it can still be proved. Terminology work in South Tyrol is necessary both for post-editing MT results and also for terminology enforcement in MT systems (e.g., through glossary functions or for LLM prompting), since the training with relatively small amounts of data leaves notable gaps.

While we know that translation errors are often related to the language pair under study, we believe that shedding light on machine translation errors for the legal language may be beneficial also for other languages. In addition, we consider that in-depth qualitative analyses are crucial for MT developers to avoid ineffective—hence environmentally unsustainable—MT training efforts and focus on other improvement strategies. A well-studied MT process, where translation scholars and computer scientists come together to tackle the main translation error types would help not only make the automatic translation process more sustainable, but also the entire translation workflow in the local administration smoother and of higher output quality.

References

Ait ElFqih, K., & Monti, J. (2023). On the Evaluation of Terminology Translation Errors in NMT and PB-SMT in the Legal Domain: A Study on the Translation of Arabic Legal Documents into English and French. *Proceedings of the First ConTenNTS Workshop and the 16th BUCC Workshop*, 26–35. https://aclanthology.org/2023.contents-1.4.pdf

Ammon, U., Bickel, H., & Lenz, A. N. (Eds.). (2016). *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen* (2nd ed.). de Gruyter.

Bane, F., Zaretskaya, A., Blanch Miró, T., Soler Uguet, C., & Torres, J. (2023). Coming to Terms with Glossary Enforcement: A Study of Three Approaches to Enforcing Terminology in

NMT. *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 345–353. https://aclanthology.org/2023.eamt-1.34.pdf

Cabrera, L., & Niehues, J. (2023). Gender Lost in Translation: How Bridging the Gap Between Languages Affects Gender Bias in Zero-Shot Multilingual Translation. In E. Vanmassenhove, B. Savoldi, L. Bentivogli, J. Daems, & J. Hackenbuchner (Eds.), *Proceedings of the 1st Workshop on Gender-Inclusive Translation Technologes* (pp. 25–35). Open Press Tilburg University. https://aclanthology.org/2023.gitt-1.3.pdf

Cao, D. (2007). *Translating Law*. Multilingual Matters.

Castilho, S., & Knowles, R. (2024). A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, 1–31. https://doi.org/doi:10.1017/nlp.2024.7

Chiocchetti, E. (2021). Effects of social evolution on terminology policy in South Tyrol. *Terminology*, *27*(1), Article 1. https://doi.org/10.1075/term.00060.chi

Chromá, M. (2008). Translating Terminology in Arbitration Discourse. In V. K. Bhatia, C. N. Candlin, J. Engberg, & J. Lung (Eds.), *Legal Discourse across Cultures and Systems* (pp. 309–328). Hong Kong University Press. https://www.jstor.org/stable/j.ctt1xwdnt.19

Contarino, A. (2021). *Neural machine translation adaptation and automatic terminology evaluation: A case study on Italian and South Tyrolean German legal texts* [Doctoral dissertation, University of Bologna]. https://amslaurea.unibo.it/24989/

Contarino, A., & De Camillis, F. (2023). Domain-adapting and evaluating machine translation for institutional German in South Tyrol. In M. Izquierdo & Z. Sanz-Villar (Eds.), *Corpus Use in Cross-linguistic Research. Paving the way for teaching, translation and professional communication* (pp. 179–194). John Benjamins. https://doi.org/10.1075/scl.113.10con

Costa-jussà, M. R. (2019). An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, *1*, 495–496. https://doi.org/10.1038/s42256-019-0105-5

De Camillis, F. (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 dissertation, University of Bologna]. http://amsdottorato.unibo.it/9695/

De Camillis, F., Stemle, E., Chiocchetti, E., & Fernicola, F. (2023). The MT@BZ corpus: Machine translation & legal language. *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 171–180. https://aclanthology.org/2023.eamt-1.17.pdf

de Groot, G.-R. (1999). Das Übersetzen juristischer Terminologie. In G.-R. de Groot & R. Schulze (Eds.), *Recht und Übersetzen* (pp. 11–46). Nomos.

Edman, L., Toral, A., & van Noord, G. (2020). Low-Resource Unsupervised NMT: Diagnosing the Problem and Providing a Linguistically Motivated Solution. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 81–90. https://aclanthology.org/2020.eamt-1.10/

Fadaee, M., Bisazza, A., & Monz, C. (2017). Data Augmentation for Low-Resource Neural Machine Translation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, 567–573. https://aclanthology.org/P17-2090.pdf

Farzindar, A., & Lapalme, G. (2009). Machine Translation of Legal Information and Its Evaluation. In Y. Gao & N. Japkowicz (Eds.), *Lecture Notes in Artificial Intelligence* (pp. 64–73). Springer. https://link.springer.com/chapter/10.1007/978-3-642-01818-3_9

Foti, M. (2022). eTranslation. Le système de traduction automatique de la Commission européenne en appui d'une Europe numérique. *Traduire*, *246*. https://doi.org/10.4000/traduire.2793

Freixa, J. (2022). Causes of terminological variation. In P. Faber & M.-C. L'Homme (Eds.), *Theoretical Perspectives on Terminology. Explaining terms, concepts and specialized knowledge* (pp. 399–420). John Benjamins. https://doi.org/10.1075/tlrp.23.18fre

Giampieri, P. (2023). *Legal Machine Translation Explained: MT in Legal Contexts*. Cambridge Scholars.

Goyle, V., Krishnaswamy, P., Ravikumar, K. G., Chattopadhyay, U., & Goyle, K. (2023). *Neural machine Translation for low resource languages*. https://aclanthology.org/2023.eamt-1.17.pdf

Haddow, B., Bawden, R., Barone, A. V. M., Helcl, J., & Birch, A. (2022). Survey of Low-Resource Machine Translation. *Computational Linguistics*, *48*(3), 673–732. https://doi.org/10.1162/coli_a_00446

Haque, R., Hasanuzzaman, M., & Way, A. (2019). Terminology Translation in Low-Resource Scenarios. *Information*, *10*(9), 273, 2–28. https://doi.org/10.3390/info10090273

Haque, R., Hasanuzzaman, M., & Way, A. (2020). Analysing terminology translation errors in statistical and neural machine translation. *Machine Translation*, *34*, 149–195. https://doi.org/10.1007/s10590-020-09251-z

Heiss, C., & Soffritti, M. (2018). DeepL Traduttore e didattica della traduzione dall'italiano in tedesco. *inTRAlinea*, *20*(1). http://www.intralinea.org/archive/article/2294

Hovy, D., Bianchi, F., & Fornaciari, T. (2020). "You Sound Just Like Your Father". Commercial Machine Translation Systems Include Stylistic Biases. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1686–1690. https://aclanthology.org/2020.acl-main.154.pdf

Ive, J., Specia, L., Szoc, S., Vanallemeersch, T., Van den Bogaert, J., Farah, E., Maroti, C., Ventura, A., & Khalilov, M. (2020). A Post-Editing Dataset in the Legal Domain: Do we Underestimate Neural Machine Translation Quality? In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, I. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the 12th Conference on Language Resources and Evaluation* (pp. 3692–3697). ELRA. https://aclanthology.org/2020.lrec-1.455.pdf

Kenny, D. (2022). Human and machine translation. In D. Kenny (Ed.), *Machine translation for everyone: Empowering users in the age of artificial intelligence* (pp. 23–50). Language Science Press. https://zenodo.org/record/6653406

Killman, J. (2014). Vocabulary Accuracy of Statistical Machine Translation in the Legal Context. In S. O'Brien, M. Simard, & L. Specia (Eds.), *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas* (pp. 85–98). Association for Machine Translation in the Americas. https://aclanthology.org/2014.amta-wptp.7/

Killman, J. (2023). Machine translation and legal terminology. Data-driven approaches to contextual accuracy. In Ł. Biel & H. J. Kockaert (Eds.), *Handbook of Terminology. Legal Terminology* (Vol. 3, pp. 485–510). Benjamins. https://benjamins.com/online/hot/articles/mac2

Kit, C., & Wong, T. M. (2008). Comparative Evaluation of Online Machine Translation Systems with Legal Texts. *Law Library Journal*, *2*(100), 299–321.

Knowles, R., Larkin, S., Tessier, M., & Simard, M. (2023). Terminology in neural machine translation: A case study of the Canadian Hansard. *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 481–488. https://nrc-publications.canada.ca/fra/voir/auteur/version/?id=808208ca-bd58-408b-b0d5-6b02f385979e

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista Tradumàtica: Tecnologies de La Traducció*, *12*, 455–463.

Martínez Domínguez, R., Rikters, M., Vasilevskis, A., Pinnis, M., & Reichenberg, P. (2020). Customized Neural Machine Translation Systems for the Swiss Legal Domain. In J. Campbell, D. Genzel, B. Huyck, & P. O'Neill-Brown (Eds.), *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas* (Vol. 2, pp. 217–223). Association for Machine Translation in the Americas. https://aclanthology.org/2020.amta-user.11.pdf

Marzi, E. (2021). La traduction automatique neuronale et les biais de genre: Le cas des noms de métiers entre l'italien et le français. *Synergies Italie*, *17*, 19–36.

Mattila, H. E. S. (2018). Legal Language. In J. Humbley, G. Budin, & C. Laurén (Eds.), *Languages for Special Purposes: An International Handbook* (pp. 113–150). De Gruyter Mouton.

Monti, J. (2020). Gender issues in machine translation. An unsolved problem? In L. von Flotow & H. Kamal (Eds.), *The Routledge handbook of translation, feminism and gender* (pp. 457–468). Routledge.

Moslem, Y., Haque, R., Kelleher, J. D., & Way, A. (2023). Adaptive Machine Translation with Large Language Models. *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 227–237. https://aclanthology.org/2023.eamt-1.22/

Moslem, Y., Romani, G., Molaei, M., Haque, R., Kelleher, J. D., & Way, A. (2023). Domain Terminology Integration into Machine Translation: Leveraging Large Language Models. *Proceedings of the Eighth Conference on Machine Translation (WMT)*, 902–911. https://aclanthology.org/2023.wmt-1.82.pdf

Mulé, M., & Johnson, C. (2010). How Effective is Machine Translation of Legal Information? *Clearinghouse Review*, *44*(1), 32–36.

Oliver, A., Alvarez, S., Stemle, E. W., & Chiocchetti, E. (2024). Training an NMT system for legal texts of a low-resource language variety (South Tyrolean German – Italian). *Proceedings of the 25th Annual Conference of the European Association for Machine Translation*, *1*, 573–579. https://eamt2024.github.io/proceedings/vol1.pdf

Pontrandolfo, G., & Quinci, C. (2023). Testing neural machine translation against different levels of specialisation. An exploratory investigation across legal genres and languages. *Trans-Kom*, *16*(1), 174–209.

Prieto Ramos, F., & Cerutti Benitez, G. (2021). Terminology as a source of difficulty in translating international legal discourses: An empirical cross-genre study. *International Journal of Legal Discourse*, *6*(2), 155–179. https://doi.org/10.1515/ijld-2021-2052

Provincial Law No. 5/2010: *Legge della Provincia autonoma di Bolzano sulla parificazione e sulla promozione delle donne e modifiche a disposizioni vigenti*: http://lexbrowser.provincia.bz.it/doc/it/lp-2010-5/legge_provinciale_8_marzo_2010_n_5.aspx

Ranathunga, S., Annie Lee, E.-S., Prifti Skenduli, M., Shekhar, R., Alam, M., & Kaur, R. (2023). *Neural Machine Translation for Low-Resource Languages: A Survey*. https://arxiv.org/abs/2106.15115

Rehm, G., & Way, A. (2023). *European Language Equality. Strategic Agenda for Digital Language Equality*. Springer.

Sánchez-Gijón, P., & Kenny, D. (2022). Selecting and preparing texts for machine translation: Pre-editing and writing for a global audience. In D. Kenny (Ed.), *Machine translation for everyone. Empowering users in the age of artificial intelligence* (pp. 81–103). Language Science Press.

Šarčević, S. (1997). *New Approach to Legal Translation*. Kluwer Law International.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender Bias in Machine Translation. In B. Roark & A. Nenkova (Eds.), *Transactions of the Association for Computational Linguistics* (Vol. 9, pp. 845–874). Association for Computational Linguistics. https://doi.org/10.1162/tacl_a_00401

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating Gender Bias in Machine Translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1679–1684. https://aclanthology.org/P19-1164.pdf

Tezcan, A., Hoste, V., & Macken, L. (2017). SCATE Taxonomy and Corpus of Machine Translation Errors. In G. Corpas Pastor & I. Durán Muñoz (Eds.), *Trends in e-tools and resources for translators and interpreters* (pp. 219–248). Brill/Rodopi. https://core.ac.uk/download/pdf/147051928.pdf

Triboulet, B., & Bouillon, P. (2023). Evaluating the Impact of Stereotypes and Language Combinations on Gender Bias Occurrence in NMT Generic Systems. In B. R. Chakravarthi, J. Griffith, K. Bali, & P. Buitelaar (Eds.), *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion* (pp. 62–70). ACL. https://aclanthology.org/2023.ltedi-1.9/

Wang, R., Tan, X., Luo, R., Qin, T., & Liu, T.-Y. (2021). A Survey on Low-Resource Neural Machine Translation. *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. https://www.ijcai.org/proceedings/2021/0629.pdf

Wiesmann, E. (2019). Machine translation in the field of law: A study of the translation of Italian legal texts into German. *Comparative Legilinguistics. International Journal for Legal Communication*, *37*, 117–153. https://doi.org/10.14746/cl.2019.37.4

Yates, S. (2006). Scaling the Tower of Babel Fish: An Analysis of the Machine Translation of Legal Information. *Law Library Journal*, *98*(3), 481–502.

Yvon, F., & Rauf, S. A. (2020). *Utilisation de ressources lexicales et terminologiques en traduction neuronale*. IMSI-CNRS. https://hal.science/hal-02895535v2