The MT@BZ corpus: machine translation & legal language

Flavia De Camillis[1] and **Egon W. Stemle**[1] and **Elena Chiocchetti**[1] and **Francesco Fernicola**[1,2]
Flavia.DeCamillis,Egon.Stemle,Elena.Chiocchetti,Francesco.Fernicola@eurac.edu
[1] Institute for Applied Linguistics, Eurac Research, Bolzano/Bozen, Italy
[2] Università di Bologna, Forlì, Italy

Abstract

The paper reports on the creation, annotation and curation of the MT@BZ corpus, a bilingual (Italian–South Tyrolean German) corpus of machine-translated legal texts from the officially multilingual Province of Bolzano, Italy. It is the first human error-annotated corpus (with an adapted SCATE taxonomy) of machine-translated legal texts in this language combination that includes a lesser-used standard variety. Project data are available on GitHub and CLARIN.[1] The output of the customized engine achieved notably better BLEU, TER and chrF2 scores than the baseline. Over 50% of the segments needed no human revision. The most frequent error categories were mistranslations and bilingual (legal) terminology errors. Our contribution brings fine-grained insights to Machine Translation Evaluation research, as it concerns a less common language combination, a lesser-used language variety and a societally relevant specialized domain. Such results are necessary to implement and inform the use of MT in institutional contexts of smaller language communities.

1 Introduction

Machine translation evaluation (MTE) assesses the performance of machine translation (MT) systems. It can be human or automatic. While automatic metrics are quickly computed and offer an idea of how a system performs, human evaluation is time-consuming and expensive but offers detailed insights into what machines get right or wrong when translating. Human MTE usually considers accuracy and fluency. Accuracy measures "the extent to which the translation transfers the meaning of the source-language unit into the target", while fluency assesses "the extent to which the translation follows the rules and norms of the target language" (Castilho et al., 2018, 18). Error classification and analysis may be considered a subtask of human MTE. It requires a detailed error taxonomy and a group of annotators (Popović, 2018, 131–32). In the past, different error taxonomies have been developed, but none was adapted or tested on the combination Italian-South Tyrolean German or on a lesser-used standard variety of a major European language.[2] A widely used error classification framework is the Multidimensional Quality Metrics (MQM) (Lommel et al., 2014), with a hierarchical list of categories and a flexible and customizable application that ensure different levels of granularity. Despite its flexibility, in our project we opted for the SCATE taxonomy, as the possibility of linking target annotations to source spans helps interpret terminology issues, our main interest. Besides, the availability of a ready-to-use annotation project with the SCATE taxonomy was an added value.

Despite the substantial improvements achieved thanks to neural technologies (Kenny, 2022, 43), MT still struggles with some language combinations in the legal domain (Wiesmann, 2019), more than in other domains (Ive et al., 2020; Foti, 2022). This is due to the inherent complexity of legal discourse, with i) terminology coming from several

[1] 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.

[2] For an overview on annotation taxonomies refer to Popović (2018) and Tezcan et al. (2017).

fields, ii) a strong relation with general language (e.g., redefined words like 'hold'), iii) convoluted syntax and long sentences and iv) abundance of internal and external references (Hiltunen, 2012; Mattila, 2018; Gotti, 2012). Legal language is particularly difficult to translate (Cao, 2007) and, consequently, to machine translate. The struggle becomes even more challenging when dealing with pluricentric languages, such as German, English or Spanish. Pluricentric languages are used in two or more countries with different official norms in grammar, orthography, and lexis (Clyne, 1991; Ammon et al., 2016). In addition, each country has a specific legal system, which is expressed by more or less diverging legal languages. This is the case of Austria, Germany, Switzerland and South Tyrol for German. In our study, we deal with South Tyrol, where the legal system in force is the Italian one and German is a co-official language at the local level.

South Tyrolean German is a standard variety of German, but no customized MT system has been developed for it so far. At the time of writing, freely accessible online MT systems have not implemented German varieties yet. In DeepL[3], two varieties are available only for English and Portuguese, while none are available in Google Translate[4]. It is reasonable to assume that most texts used to train MT engines in German come from the European Union and Germany. Their performance on South Tyrolean German necessarily fails to consider typical local terminology and phraseology, as already proved by Heiss and Soffritti (2018).

Against this background, we identified two major research gaps we aim to contribute to. To our knowledge, no corpus of machine translated legal texts has been annotated so far, nor does a corpus for the combination Italian-South Tyrolean German exist. The MT@BZ corpus intends to: i) contribute to the evaluation of machine translated legal language; ii) set the basis for creating an MT system for South Tyrolean public institutions.

2 Motivation

The South Tyrolean local administration is required to publish laws, decrees, circulars, communications, etc. in both Italian and German (Presidential Decree 670/1972). This is done by translating them from either language. The task is usually carried out by civil servants, generally non-professional translators (De Camillis, 2021). They use also freely accessible online MT systems, which underperfom in the South Tyrolean legal language (see Section 1). To address the research gap related to customized MT for South Tyrol, we considered an annotated corpus of MT errors a useful first step for the following reasons. First, no annotation scheme has been tested on the combination Italian-German and no annotated corpus exists for this language pair. We consider it of utmost importance to assess the performance of MT systems on less common language combinations to identify language-dependent issues more clearly. Furthermore, our research scenario deals with a lesser-used standard variety (South Tyrolean German), for which the development of specific language technologies was hardly addressed so far. Finally, legal language is an essential aspect of implementing linguistic human rights pertaining to language minorities, such as the right to understand the language within court proceedings (Skutnabb-Kangas, 2012).

Second, not many scholars working on MTE focused on legal language. Among those who did (Wiesmann, 2019; Ive et al., 2020; Farzindar and Lapalme, 2009; Yates, 2006; Kit and Wong, 2008; Mulé and Johnson, 2010), only Farzindar and Lapalme assessed the quality of Canadian court judgements translated between English and French with three human evaluators. However, they did not annotate MT mistakes. No fully annotated corpus of machine-translated legal texts has been created so far, to the best of our knowledge.

Third, a fine-grained error annotation could be used to advance research in the field of Quality Estimation (QE). The latest WMT shared task adopted MQM to produce the human gold standard for the task datasets (Zerva et al., 2022) because fine-grained annotation schemas are more reliable for the metrics task (Freitag et al., 2021).

Last, an annotated corpus is a valuable research output and may serve as input for further research. It sheds light on the mistakes of an MT system when translating legal texts in a given language pair, thus contributing to MT research on legal language. It can also represent useful input to further develop or fine-tune a customized MT system exploiting high-quality human, granular and refined analyses.

[3]https://www.deepl.com/translator
[4]https://translate.google.com/

	documents		tokens	
	IT	DE	IT	DE
General	10	10	24,300	20,339
COVID-19	16	16	14,506	12,663
Total	26	26	38,806	33,002

Table 1: Overview on the texts in the MT@BZ corpus

3 Data compiling

The MT@BZ corpus was compiled in late 2020 by downloading a set of provincial decrees from the local legal database LexBrowser in both Italian and German.[5] We selected a range of decrees published from November 2020. In October 2020, Contarino (2021) created a bilingual aligned corpus of texts from LexBrowser (LEXB), which we used to train an MT system that translated the MT@BZ corpus. It was therefore essential that the decrees of our corpus were not included in LEXB. The aim of our test was to assess the performance of a customized MT system in a "real world" scenario.

3.1 Data selection

To assess potential differences in the performance of the engine, we selected 26 decrees covering an array of topics (education, insurance, construction, COVID-19; etc). We excluded very short decrees and decrees consisting mostly of tables, as we wanted to evaluate the performance on running text and enough context span. The average length of the decrees is 1,400 tokens. We also preferred decrees related to topics covered by the local terminological resource, bistro[6,7]. In total, we collected 52 texts, 26 in Italian and the corresponding 26 in German. The overall amount of tokens is 72,000 (see Table 1 for more details). The decrees were downloaded in PDF-format, converted to TXT by hand and aligned. The alignment was then polished manually.

3.2 Data translation

In our translation scenario, 26 texts were translated from Italian into South Tyrolean German and 26 from South Tyrolean German into Italian using ModernMT (MMT) (Germann et al., 2016).[8] to follow up on previous tests (Contarino, 2021; De Camillis, 2021).

MMT is based on the state-of-the-art Transformer architecture. It is trained on a large pool of parallel data and employs an instance-based adaptation approach described by Farajian et al. (2017). It requires a baseline model, an in-domain adaption corpus and a segment to be translated. A set of source-target sentence pairs is retrieved, whose source is similar to the given segment. With this data, the parameters of the neural network model are locally fine-tuned before translation. After having translated the sentence, the adapted model is reset to the parameters of the original system. Such an approach has shown significant improvements in the translation of terminology (Farajian et al., 2018). Another reason for choosing MMT resides in the easiness of customization, since the user only has to upload one or more translation memories that train the basic engine.

We exploited the plug-in of MMT in our usual translation environment RWS Trados Studio. In October 2021, after having created two different projects (IT>DE, DE>IT), we first translated the texts using the default memory available in MMT, MyMemory, which we consider our baseline. Then, we repeated the process by uploading the LEXB corpus into MMT together with some extra material: 20 national laws (with their official translations into German) and some small translation memories from the local Office for Language Issues.[9] The uploaded memory had 230,402 bilingual segments (but only 202,779 after the conversion in RWS Trados Studio). The memory had previously been accurately cleaned, excluding very long and very short sentences, identical or almost-identical segments, corrupted segments, segments with wrong source or target language, etc. The cleaning process applied the scripts by Contarino (2021). Finally, we translated the texts using the MT function in Trados Studio and exported the files in TXT. The output of the customized engine achieved a higher level of quality over the baseline according to the automatic scores BLEU, TER and chrF2 (see Table 2). If we exclude perfect matches (for further details see Section 5.1), we can still see an improvement according to all three scores (see Table 3).

[5] http://lexbrowser.provinz.bz.it/.

[6] https://bistro.eurac.edu/.

[7] We uploaded a translation memory into ModernMT that contained an export of source and target terms (term-to-term segments). However, this step did not influence results, possibly because neural MT learns terms within a given context rather than from lists.

[8] https://www.modernmt.com/.

[9] The Office for Language Issues is the only translation office within South Tyrol's provincial administration. They agreed to share their TMs with us for research purposes.

	BLEU	TER	chrF2
DE-IT			
Baseline	26.65	66.86	52.97
Customized	71.22	23.14	84.43
IT-DE			
Baseline	27.59	64.21	55.60
Customized	74.74	23.72	84.27

Table 2: BLEU, TER and chrF2 scores for DE>IT and IT>DE sub-corpora of the MT@BZ corpus

	BLEU	TER	chrF2
DE-IT			
Baseline	25.49	69.44	51.64
Customized	51.95	41.10	71.96
IT-DE			
Baseline	27.11	66.26	54.59
Customized	50.78	45.88	69.18

Table 3: BLEU, TER and chrF2 scores for DE>IT and IT>DE sub-corpora of the MT@BZ corpus excluding perfect matches

3.3 Data outlook

Overall, we have 104 texts: A) 26 source texts in Italian (that serve as reference translations for the corresponding decrees in German)[10]; B) 26 source texts in German (also reference translations); C) 26 baseline machine translations in Italian and German respectively; D) 26 customized machine translations in Italian and German respectively.

In other words, for each text there is: i) a source text, ii) a reference (human) translation, iii) a translation done by baseline MMT, iv) a translation done by customized MMT. We have shared all texts with the research community. [1]

4 Annotation

We annotated our corpus to identify the more frequent error categories produced by a customized 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 discourse. Among the many available, we selected the SCATE taxonomy (Tezcan et al., 2017) for several reasons. It has been used in a similar annotation campaign and, as such, allows for accuracy and fluency errors annotations. It allows to link accuracy errors in the target to relevant spans in the source language. It is provided with detailed guidelines. It is detailed but not to an unsustainable level. Finally, it is easy to implement, as an annotation project carried out by the SCATE group was shared as a complete WebAnno project with the research community (Fonteyne et al., 2020).

4.1 Scheme development

The SCATE taxonomy was originally developed for the language combination English-Dutch[11]. The guidelines come with a great number of examples. We kept the basic structure of the guidelines (version 1.3.3)[12] and used English to facilitate comparability, while we adapted the examples and some categories to our use-case. The major changes consisted in: i) adding the Accuracy category "Gender"; ii) excluding the fourth level subcategories for Word-sense disambiguation; iii) adding the Fluency category "Coherence";[13] iv) excluding the fourth level subcategories for Word form, Extra words, Lexical choice, Spelling.

The additions were necessary because we identified two new error categories while testing the guidelines. The exclusions are due to technical challenges of a user-friendly implementation. Even though WebAnno allows for a fourth level of annotation, user interaction for annotations on this level is cumbersome.

Adapting the guidelines required long considerations as to the selection of examples from the MT@BZ corpus. For some categories, it was impossible to find in-project examples because the mistake did not occur in our corpus. This relates to some fluency categories and might depend on the fact that neural MT usually makes less language/formal mistakes[14].

4.2 Scheme design

The annotation scheme is divided in two sections, as shown in Figure 1: *Accuracy* and *Fluency*. Accuracy errors concern the transfer of meaning from

[10] It is not possible to determine the source language for legal texts published in the multilingual setting of the Province of Bolzano: text drafting may occur in more than one language and extracts of published texts may be reused in either language. For this reason, we considered both versions of each decree either source or reference translation in the corresponding test settings.

[11] https://users.ugent.be/~atezcan/.

[12] There are minor discrepancies between the taxonomy in the guidelines and the taxonomy in Tezcan et al. (2017). We adapted our taxonomy starting from the guidelines.

[13] The category "Co-reference" is as greyed out in Figure 1, because we excluded it during the campaign (see Section 5.2).

[14] The original guidelines were published in 2015. Neural technologies were released soon after.

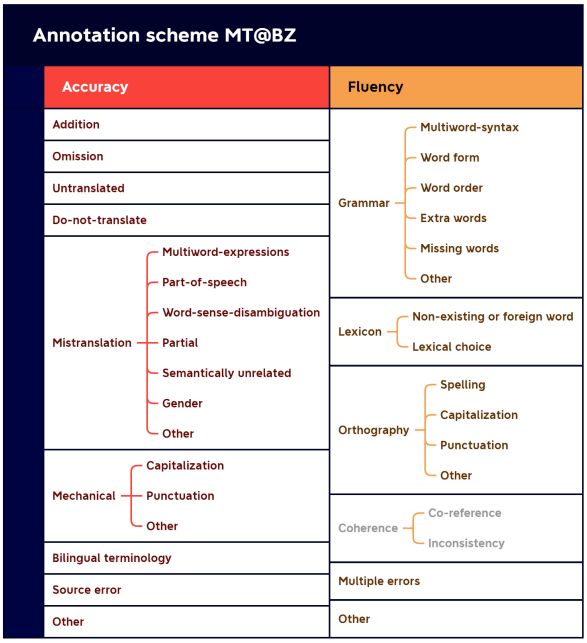

Figure 1: Annotation scheme of the MT@BZ Corpus

source to target and are therefore 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 two sub-levels.

4.3 Annotation examples

Figure 2 shows a simple annotation on a short segment with only one mistake. Figure 3 shows more complex annotations, with several mistakes identified. The majority of mistakes are mistranslations, where the sense was misinterpreted (Word-sense disambiguation, Semantically unrelated), and errors relating to South Tyrolean legal terminology (Bilingual terminology).

5 Data preparation and annotation workflow

The annotation campaign was carried out between June and November 2022 with the help of four annotators with a degree in translation. Two are Italian native speakers, one is a native speaker of (South Tyrolean) German, one is a balanced bilingual. All have at least a C1 level in their second language. One Italian native speaker annotated only six texts. The three other annotators worked on both language directions due to a shortage of annotators. One is a Master's student in translation, one has more than 5 years of experience and one has 20 years of experience in translation. Two translators, one with 20+ years and one with 5+ years, were entrusted with the final curation step, which they performed on texts translated into their native language.

5.1 Data preparation

The original data is maintained in Excel files, with individual lines corresponding to text segments of the original text. The columns contain: 1) source segment, 2) human reference translation, 3) baseline MMT output, 4) customised MMT output.

For further processing in WebAnno, we converted the data into the WebAnno TSV 3.3 File Format[15] with our own script.[16] This yielded files where corresponding segments are paired: the source segment is paired with the customised MMT output in one line with a line break between the segments. If the customized MMT output text segment is (almost)[17] identical to the human reference translation, the whole segment is "pre-annotated" as a "perfect match". This relates to human reference translations having been re-used from the translation memory and that should be disregarded during annotation. However, we did not exclude these segments from the data to keep the context for the following text segments.

To be able to annotate complete words independently, even in case of incorrect separation (tokenization) from the surrounding characters, we used the NLTK (Bird et al., 2009) *nltk.tokenize.regexp.WordPunctTokenizer*, which tokenizes a text into a sequence of alphabetic and non-alphabetic characters. In this way, annotators had easy access to words and individual characters during the annotation campaign.

We finally loaded the available StylesNMT project into our local WebAnno installation, deleted their file set, upload ours and adapted the annotation layer and tagset settings to our needs. Overall, having a readily available project to start from made the task easier. During the course of the project and due to technical reasons, we flawlessly switched to INCEpTION.

5.2 Limitations

We identified four major limitations of our project. First, we did not include style errors, unlike other

[15] https://webanno.github.io/webanno/releases/3.6.11/docs/user-guide.html#sect_webannotsv.

[16] Available at https://gitlab.inf.unibz.it/commul/mt-bz.

[17] To be a "perfect match" two segments must be identical, except for the occurrence of these special characters: <", ’, ´, ‘, ’, ‘, U+201B, „, U+201F, “, ”, —, –, °, «, », ‹, ›, . . . >.

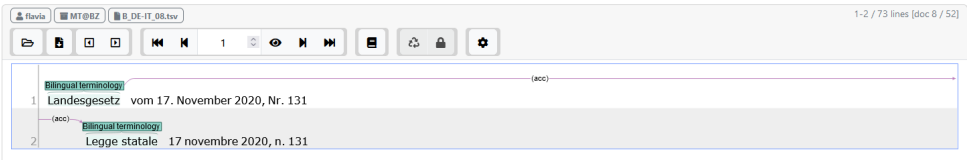

Figure 2: Example of a simple annotation

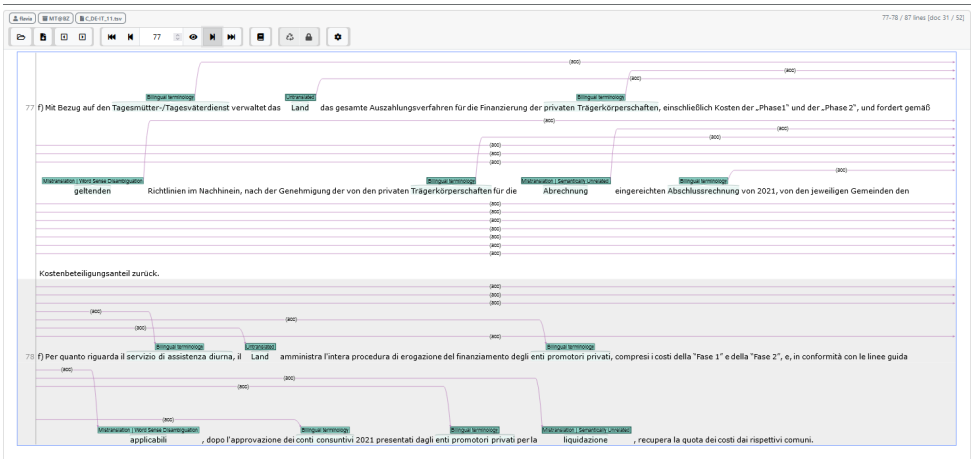

Figure 3: Example of a complex annotation

taxonomies, such as MQM (Lommel et al., 2014) and the SCATE taxonomy in a later version (Tezcan et al., 2019), as we considered them less relevant for our text type. Second, we did not use scalar metrics nor questionnaires to assess translation quality, as others did (e.g., Freitag et al. (2021); Castilho (2021)), mainly because we were mostly interested in classifying errors. Third, we annotated at segment-level rather than at document-level, even though some issues (e.g., gender or coherence errors) would have been better annotated at document-level. Last, we only used one MT system, which does not allow for generalisations. However, we share both the corpus and the guidelines with the community, so that replication studies can be carried out and results compared.

5.3 Annotation

Prior to annotation, two annotators tested the guidelines extensively to select the most adequate examples from the corpus and discuss overlaps between categories. All four annotators checked their work after the first round of annotations. The overall amount of hours spent on this task is around one person-month each.

The annotation process made some key aspects evident. The most striking result is that over 50% of the segments needed no human revision, as they were identical to the reference text. This is an impressive result, if we consider the potential use of translation memories and of an MT system trained with these memories in a public institution where civil servants translate on a daily basis: notable amounts of text could be re-used from past publications.

The most represented error categories in the other segments were Accuracy mistakes, mainly Mistranslations and Bilingual terminology mistakes. Mistranslations are related to sense. Homonymy, terminology from different domains and context-related nuances are typical elements of legal discourse and usually hard to disambiguate for a machine. Bilingual terminology errors include translations of legal terms with a general language equivalent and translations that could be considered correct within another legal system but do not correspond to local legal terminology (e.g., *Paragraph* is a subdivision of legal texts used in German law but South Tyrolean legislation uses *Artikel*). Despite careful redefinition of the Bilingual terminology category according to the South Tyrolean terminological standards, in many occasions annotators disagreed as to whether a mistake was to be classified as Bilingual terminology, Word-sense disambiguation or Semantically unrelated.

Multiword-expressions was a further frequent error category. It relates to a typical feature of legal discourse, i.e., titles of legal texts and legal

	Annotator 1		Annotator 2		Annotator 3	
Addition	144	0.04	143	0.04	192	0.07
Omission	322	0.09	225	0.07	313	0.12
Untranslated	36	0.01	23	0.01	32	0.01
Do-not-translate	0	0.00	0	0.00	0	0.00
Mistranslation	1789	0.49	1671	0.50	1527	0.58
Mechanical	130	0.04	90	0.03	134	0.05
Bilingual terminology	1146	0.32	1129	0.34	433	0.16
Source error	29	0.01	6	0.00	4	0.00
Other	21	0.01	35	0.01	2	0.00
	3617		3322		2637	

Table 4: Number of annotations per annotator (% of annotations)

	Annotator 1		Annotator 2		Annotator 3	
Grammar	267	0.32	198	0.43	152	0.52
Lexicon	124	0.15	54	0.12	2	0.01
Orthography	275	0.33	208	0.45	137	0.47
Coherence	150	0.18	2	0.00	0	0.00
Multiple errors	1	0.00	1	0.00	0	0.00
Other	29	0.03	1	0.00	0	0.00
	846		464		291	

Table 5: Number of annotations per annotator (% of annotations)

phraseology. Titles were rarely reproduced in their correct wording but translated *ex novo*. Phraseology was often translated literally.

Finally, Fluency mistakes were generally less frequent. Missing words, Word order, Punctuation and Spelling mistakes (the latter only for German) were the most recurrent ones. Morphology was most of the times correct, with the exception of diacritics and punctuation. Punctuation errors occurred more frequently in long bullet point segments. See Tables 4 and 5 for details.

5.4 Inter-Annotator-Agreement

Annotators usually face two tasks: they must locate an error and assign it to one or more error categories. This means that annotations can differ in two ways: 1) the location and span of an error (i.e., over which words or characters it spreads) and 2) the type of error identified. Inter-Annotator-Agreement (IAA) is a method to assess to what extent the annotators agree with each other and the reliability of their annotations. Much work has been done towards assessing the situation when the segments to be annotated are known (Artstein, 2017) but very few methods are proposed and discussed for the joint tasks of locating segments and labeling them.

The method we used for IAA calculations is the Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment (Mathet et al., 2015), which unifies the process of measuring alignment and agreement. [18,19] Similar to kappa-agreement (κ), γ-agreement is a chance-adjusted measure of the agreement between annotators.

The overall gamma-value for all Accuracy annotations is 0.73 and for all Fluency annotations it is 0.77, which can be considered a good level of agreement. See Tables 6 and 7 for details. [20]

5.5 Gold standard

Two annotators, native speakers of the respective target language with many years of experience, curated the gold standard for each translation direction. The decision not to subdivide work in other ways (e.g., by subgroup of texts) aimed at achieving a possibly high consistency within a specific translation direction.

[18]For the actual calculations, we use INCEpTALYTICS: https://doi.org/10.5281/zenodo.7095346

[19]IAA was calculated on the performance of the three annotators who completed the annotation task.

[20]Note that since annotation spans may overlap, the mean of the individual values from the tables is different from those given here.

	γ-value
Addition	0.94
Omission	0.86
Untranslated	0.97
Do-not-translate	
Mistranslation	0.74
Mechanical	0.94
Bilingual terminology	0.82
Source error	1
Other	0.68

Table 6: Individual gamma-values per Accuracy category

	γ-value
Grammar	0.88
Lexicon	0.32
Orthography	0.94
Coherence	0.00
Multiple errors	0.00
Other	0.00

Table 7: Individual gamma-values per Fluency category

Curation was essential since diverging annotations related to error types or spans were frequent in our data set. This was due to i) human errors, ii) varying views on what an "error" is but more often to iii) different interpretations of the guidelines. In the latter case, curation becomes a useful step to fine-tune annotations guidelines for future campaigns.

Different annotations due to human mistakes included the circumstances where one or more annotators overlooked an error, accidentally selected the wrong error category or forgot to indicate a more-fine-grained annotation where available. It also happened that an annotator considered a mistake what other annotators rather classified as an imperfection not worthy of being annotated.

More frequently, diverging annotations were due to different interpretations of the guidelines or to insufficient information shared via the guidelines. This affected both annotation spans and error annotations. Inconsistencies as to annotated spans mainly concerned articles and punctuation. The decision to include or exclude articles in some error annotations (e.g., Bilingual terminology, Gender) was a frequent cause of diverging annotations. Punctuation (e.g., commas, full stops) also tended to be deliberately excluded from error annotation by some annotators while others did not pay systematic attention. Span inconsistencies related to a different interpretation of the guidelines concerned Gender errors. For example, one annotator systematically and consistently annotated the omitted part of male and female couplets rather than the entire span. Another frequent difference concerned complex terms that contained other terms (e.g., *Dekret des Landeshauptmanns*, decree of the president of the province). With mistakes happening often at sub-term level, some annotators marked only a part of the complex term (*Landeshauptmann*), others the entire term. To keep annotation as elementary as possible, during curation the first choice was considered more appropriate and applied throughout the curated data set.

The curators had to resolve annotation inconsistencies for the three error categories that were more likely to be interpreted differently, i.e. Bilingual terminology (BT), Semantically unrelated (SU) and Word-sense disambiguation (WSD). To adopt a clear line, any term related to the Italian or local legal system and administration, especially if present in *bistro*, was considered a BT error. Whenever it was possible to translate the source term with the given target term in some contexts, it was considered WSD. Contrarily, when it was impossible to translate a given source term with a given target term, it was considered a SU error.

The error categories Other under Mistranslation and Accuracy posed particular challenges, especially when the MT system could not interpret the references between the words in the source texts correctly. In more complex cases, diverging annotations were plausible and sensible, so that the curator had to follow a possibly consistent line throughout the entire set of texts.

6 Conclusion

We reported the creation, annotation and curation of the corpus MT@BZ, a bilingual (Italian–South Tyrolean German) corpus of machine translated legal texts from the Province of Bolzano. To the best of our knowledge, this is the first annotated corpus of machine translated texts from the legal domain for a combination of languages that also includes a lesser-used standard language variety. It includes 52 decrees (26 in Italian and the corresponding 26 in South Tyrolean German) for an overall amount of 72,000 tokens. We selected and retrieved the texts from the institutional pages of the local administration of South Tyrol and translated them with the help of the MMT engine plugged-in in the RWS Studio environment. A baseline translation was acquired with the default translation memory integrated in MMT, while a customized output came from the integration of a 230,000 segments trans-

lation memory of bilingual legislation. The customized engine outperformed the baseline according to BLEU, TER and chrF2 scores. We annotated translation errors on the customized machine translation outputs, using the SCATE taxonomy (Tezcan et al., 2017) adapted to our language pair. Three annotators annotated the entire corpus achieving a good level of agreement (IAA 0,74 - gamma-value). Finally, we curated the corpus to produce a gold standard.

We believe our contribution brings more fine-grained insights to the field of Machine Translation Evaluation, because we considered both a lesser-common language combination, a lesser-common language variety and a specialized domain, even if we focused exclusively on error classification and used only one MT system, which does not allow for generalisations. This kind of very granular evaluations seems necessary to integrate the use of MT in institutional contexts of smaller realities like South Tyrol. We have shared[1] the corpus and the guidelines, as well as the project data, with the community to foster replication studies but also to encourage MT researchers to focus on lesser-used languages in real life scenarios, such as public institutions in minority language communities.

References

Ammon, Ulrich, Hans Bickel, and Alexandra N. Lenz, editors. 2016. *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen*. de Gruyter, Berlin, 2nd edition.

Artstein, Ron, 2017. *Handbook of Linguistic Annotation*, chapter Inter-annotator Agreement, pages 297–313. Springer Netherlands, Dordrecht.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc.

Cao, Deborah. 2007. *Translating Law*. Multilingual Matters, Clevedon.

Castilho, Sheila, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. Approaches to Human and Machine Translation Quality Assessment. In Moorkens, Joss, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors, *Translation Quality Assessment: From Principles to Practice*, Machine Translation: Technologies and Applications, pages 9–38. Springer International Publishing, Cham.

Castilho, Sheila. 2021. Towards Document-Level Human MT Evaluation: On the Issues of Annotator Agreement, Effort and Misevaluation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 34–45. Association for Computational Linguistics.

Clyne, Michael. 1991. Pluricentric Languages – Introduction. In Clyne, Michael, editor, *Pluricentric Languages: Differing Norms in Different Nations*, pages 1–10. De Gruyter Mouton, Berlin, Boston.

Contarino, Antonio. 2021. Neural Machine Translation Adaptation and Automatic Terminology Evaluation: A Case Study on Italian and South Tyrolean German Legal Texts. Master's thesis, Università di Bologna, Bologna, Italy.

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*. Ph.D. thesis, Alma Mater Studiorum - Università di Bologna.

Farajian, M. Amin, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137, Copenhagen, Denmark. Association for Computational Linguistics.

Farajian, M Amin, Nicola Bertoldi, Matteo Negri, Marco Turchi, and Marcello Federico. 2018. Evaluation of Terminology Translation in Instance-Based Neural MT Adaptation. In Pérez-Ortiz, Juan Antonio, Felipe Sánchez-Martínez, Miquel Esplà-Gomis, Maja Popović, Celia Rico, André Martins, Joachim Van den Bogaert, and Mikel L. Forcada, editors, *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 149–158, Alacant, Spain. Universitat d'Alacant.

Farzindar, Atefeh and Guy Lapalme. 2009. Machine Translation of Legal Information and Its Evaluation. In Gao, Yong and Nathalie Japkowicz, editors, *Advances in Artificial Intelligence*, Lecture Notes in Computer Science, pages 64–73, Berlin, Heidelberg. Springer.

Fonteyne, Margot, Arda Tezcan, and Lieve Macken. 2020. Literary Machine Translation under the Magnifying Glass: Assessing the Quality of an NMT-Translated Detective Novel on Document Level. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3790–3798, Marseille, France. European Language Resources Association.

Foti, Markus. 2022. eTranslation. Le système de traduction automatique de la Commission européenne en appui d'une Europe numérique. *Traduire*, 246:28–35.

Freitag, Markus, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Germann, U., E. Barbu, L. Bentivogli, N. Bertoldi, N. Bogoychev, C. Buck, D. Caroselli, L. Carvalho, A. Cattelan, R. Cettolo, M. Federico, B. Haddow, D. Madl, L. Mastrostefano, P. Mathur, A. Ruopp, A. Samiotou, V. Sudharshan, M. Trombetti, and J. van der Meer. 2016. Modern MT: a new open-source machine translation platform for the translation industry. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, Riga, Latvia. Baltic Journal of Modern Computing.

Gotti, Maurizio. 2012. Text And Genre. In Tiersma, Peter M. and Lawrence Solan, editors, *The Oxford Handbook of Language and Law*, pages 52–66. Oxford University Press, Oxford.

Heiss, Christine and Marcello Soffritti. 2018. DeepL Traduttore e didattica della traduzione dall'italiano in tedesco - Alcune valutazioni preliminari. *inTRAlinea online translation journal*, 20:1–11.

Hiltunen, Risto. 2012. The Grammar And Structure Of Legal Texts. In Solan, Lawrence M. and Peter M. Tiersma, editors, *The Oxford Handbook of Language and Law*, pages 39–51. Oxford University Press, Oxford.

Ive, Julia, Lucia Specia, Sara Szoc, Tom Vanallemeersch, Joachim Van den Bogaert, Eduardo Farah, Christine Maroti, Artur Ventura, and Maxim Khalilov. 2020. A Post-Editing Dataset in the Legal Domain: Do we Underestimate Neural Machine Translation Quality? In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 3692–3697. European Language Resources Association (ELRA), Marseille.

Kenny, Dorothy. 2022. Human and machine translation. In Kenny, Dorothy, editor, *Machine translation for everyone: Empowering users in the age of artificial intelligence*, number 18 in Translation and Multilingual Natural Language Processing. Language Science Press, Berlin.

Kit, Chunyu and Tak Ming Wong. 2008. Comparative Evaluation of Online Machine Translation Systems with Legal Texts. *Law Library Journal*.

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

Mathet, Yann, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Mattila, Heikki E.S. 2018. Legal Language. In Humbley, John, Gerhard Budin, and Christer Laurén, editors, *Languages for Special Purposes: An International Handbook*, pages 113–150. De Gruyter Mouton, Berlin, Boston.

Mulé, Michael and Claudia Johnson. 2010. How effective is machine translation of legal information? In *Clearinghouse Review Journal of Poverty Law and Policy*, volume 44. Shriver Center on Poverty Law, thirty-second edition.

Popović, Maja. 2018. Error classification and analysis for machine translation quality assessment. *Translation quality assessment: From principles to practice*, pages 129–158.

Skutnabb-Kangas, Tove. 2012. Linguistic Human Rights. In Solan, Lawrence M. and Peter M. Tiersma, editors, *The Oxford Handbook of Language and Law*, pages 235–247. Oxford University Press, Oxford.

Tezcan, Arda, Véronique Hoste, and Lieve Macken. 2017. SCATE Taxonomy and Corpus of Machine Translation Errors. In Corpas Pastor, Gloria and Isabel Duran-Munoz, editors, *Trends in e-tools and resources for translators and interpreters*, volume 45 of *Approaches to translation studies*, pages 219–248. Brill-Rodopi, Leiden and Boston.

Tezcan, Arda, Joke Daems, and Lieve Macken. 2019. When a 'sport' is a person and other issues for NMT of novels. In *Proceedings of the Qualities of Literary Machine Translation*, pages 40–49, Dublin, Ireland. European Association for Machine Translation.

Wiesmann, Eva. 2019. Machine translation in the field of law: a study of the translation of Italian legal texts into German. *Comparative Legilinguistics*, 37:117–153.

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

Zerva, Chrysoula, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 Shared Task on Quality Estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.