SHARING HIGH-QUALITY LANGUAGE RESOURCES IN THE LEGAL DOMAIN TO DEVELOP NEURAL MACHINE TRANSLATION FOR UNDER-RESOURCED EUROPEAN LANGUAGES

Petra Bago, Sheila Castilho, Edoardo Celeste, Jane Dunne, Federico Gaspari, Níels Rúnar Gíslason, Andre Kåsen, Filip Klubička, Gauti Kristmannsson, Helen McHugh, Róisín Moran, Órla Ní Loinsigh, Jon Arild Olsen, Carla Parra Escartín, Akshai Ramesh, Natalia Resende, Páraic Sheridan, Andy Way*

Abstract

This article reports some of the main achievements of the European Union-funded PRINCIPLE project in collecting high-quality language resources (LRs) in the legal domain for four under-resourced European languages: Croatian, Irish, Norwegian, and Icelandic. After illustrating the significance of this work for developing translation technologies in the context of the European Union and the European Economic Area, the article outlines the main steps of data collection, curation, and sharing of the LRs gathered with the support of public and private data contributors. This is followed by a description of the development pipeline and key features of the state-of-the-art, bespoke neural machine translation (MT) engines for the legal domain that were built using this data. The MT systems were evaluated with a combination of automatic and human methods to validate the quality of the LRs collected in the project, and the high-quality LRs were subsequently shared with the wider community via the ELRC-SHARE repository. The main challenges encountered in this work are discussed, emphasising the importance and the key benefits of sharing high-quality digital LRs.

Keywords: language resources; under-resourced languages; legal translation; neural machine translation; evaluation.

COMPARTIR RECURSOS LINGÜÍSTICS DE QUALITAT EN L'ÀMBIT JURÍDIC PER DESENVOLUPAR LA TRADUCCIÓ AUTOMÀTICA NEURONAL PER A LES LLENGÜES EUROPEES AMB POCS RECURSOS

Resum

En aquest article es presenten algunes de les reeixides principals del projecte PRINCIPLE, finançat per la Unió Europea, en relació amb la recopilació de recursos lingüístics (RL) de qualitat en l'àmbit jurídic per a quatre llengües europees amb pocs recursos: el croat, l'irlandès, el noruec i l'islandès. Després d'il·lustrar la importància d'aquest treball per desenvolupar tecnologies de la traducció en el marc de la Unió Europea i l'Espai Econòmic Europeu, en l'article es descriuen els passos principals per recollir dades, conservar i compartir els RL recopilats amb l'ajuda de proveïdors de dades públics i privats. A continuació, es descriuen el procés de desenvolupament i les característiques clau dels motors de traducció automàtica (TA) neuronal d'última generació adaptats al context jurídic que s'han creat amb aquestes dades. Els sistemes de TA es van avaluar amb una combinació de mètodes automàtics i humans per validar la qualitat dels RL recollits en el projecte i, més endavant, els RL de qualitat es van compartir amb el públic general a través del repositori ELRC-SHARE. L'article debat les dificultats més importants amb què ha topat aquest treball i s'hi subratlla la importància i els avantatges primordials de compartir RL digitals de qualitat.

Paraules clau: recursos lingüístics; llengües amb pocs recursos; traducció jurídica; traducció automàtica neuronal; avaluació.

* Petra Bago, Faculty of Humanities and Social Sciences, University of Zagreb – Croatia, pbago@ffzg.hr. Sheila Castilho, ADAPT Centre, Dublin City University, sheila.castilho@adaptcentre.ie. Edoardo Celeste, School of Law and Government, Dublin City University and ADAPT Centre – Ireland, edoardo.celeste@adaptcentre.ie. Jane Dunne, ADAPT Centre, Dublin City University, jane.dunne@adaptcentre.ie. Federico Gaspari, ADAPT Centre, Dublin City University, federico.gaspari@adaptcentre.ie. Níels Rúnar Gíslason, University of Iceland, niels@hi.is. Andre Kåsen, National Library of Norway, andre.kasen@nb.no. Filip Klubička, Faculty of Humanities and Social Sciences, University of Zagreb – Croatia, filip.klubicka@adaptcentre.ie. Gauti Kristmannsson, University of Iceland, gautikri@hi.is. Helen McHugh, ADAPT Centre, Dublin City University, helen.mchugh@adaptcentre.ie. Róisín Moran, Iconic Translation Machines Ltd. – Ireland, roisin@iconictranslation.com. Órla Ní Loinsigh, ADAPT Centre, Dublin City University, orla.niloinsigh@adaptcentre.ie. Jon Arild Olsen, National Library of Norway, jon.olsen@nb.no. Carla Parra Escartín, Iconic Translation Machines Ltd. – Ireland, carla@iconictranslation.com. Akshai Ramesh, Iconic Translation Machines Ltd. – Ireland, akshai@iconictranslation.com. Natalia Resende, ADAPT Centre, Dublin City University – Ireland, natalia.resende@adaptcentre.ie. Páraic Sheridan, Iconic Translation Machines Ltd. – Ireland, paraic@iconictranslation.com. Andy Way, ADAPT Centre, Dublin City University, andy.way@adaptcentre.ie

Article received: 29.10.2021. Blind reviews: 11.01.2022 and 18.02.2022. Final version accepted: 29.09.2022.

Recommended citation: Petra Bago, Sheila Castilho, Edoardo Celeste, Jane Dunne, Federico Gaspari, Níels Rúnar Gíslason, Andre Kåsen, Filip Klubička, Gauti Kristmannsson, Helen McHugh, Róisín Moran, Órla Ní Loinsigh, Jon Arild Olsen, Carla Parra Escartín, Akshai Ramesh, Natalia Resende, Páraic Sheridan, Andy Way. (2022). Sharing high-quality language resources in the legal domain to develop neural machine translation for under-resourced European languages. *Revista de Llengua i Dret, Journal of Language and Law*, *78*, 9-34. https://www.doi.org/10.2436/rld.i78.2022.3741

Contents

1 Introduction and background

Europe has a rich and complex linguistic fabric (Gaspari et al., 2022, p. 2). While preserving and promoting this diverse heritage is one of the stated aims of the European Union (EU), this endeavour presents several challenges. Upholding the equality of all languages in the current digital age is especially difficult for those with relatively small populations of speakers that have not traditionally been well supported in terms of technologies and tools such as machine translation (MT), speech synthesis, speech recognition, etc. Against this background, this article presents some of the main achievements of the EU-funded PRINCIPLE[1] project to support four under-resourced European languages – Croatian, Irish, Norwegian,[2] and Icelandic – with a view to ensuring their speakers benefit from the best possible language technologies and tools.[3] According to recent data published in a collection of comprehensive language reports authored by leading experts of these languages as part of the EU-funded European Language Equality project, Croatian has over 5.5 million speakers (Tadić, 2022, p. 3), Irish about 1.7 million (Lynn, 2022, p. 3), Norwegian roughly 5 million (Eide et al., 2022, p. 4), and Icelandic approximately 370,000 (Rögnvaldsson, 2022, p. 4). Therefore, the PRINCIPLE project had the potential to directly benefit well over 12.5 million citizens overall. In addition, considering that some of the MT engines developed in the project translate from these languages into English, the global population that could potentially benefit is much larger.

The PRINCIPLE project focused on collecting high-quality digital language resources (LRs) in selected high-priority areas for the EU. Its previous progress was reported in Way and Gaspari (2019), Bago et al. (2020), and Bago et al. (2022), and in this article, we cover specifically the legal domain. An important aspect of PRINCIPLE was the use of the LRs collected to develop state-of-the-art, bespoke domain-specific neural MT (NMT) engines, which were evaluated to validate the quality of the LRs, before the data sets were made publicly available via the ELRC-SHARE repository.[4] In addition, these digital LRs were passed on to the European Commission to improve eTranslation,[5] the online MT engine freely available to staff working for EU institutions and agencies, as well as to employees of public administrations, small- and medium-sized enterprises, and university language faculty in any EU Member State, Iceland, and Norway. Extensive, high-quality LRs are crucial to improving the performance of eTranslation, especially in the translation of specialised documents. In this article, we take the legal domain as a case in point to demonstrate the importance of collecting and sharing high-quality LRs, especially for under-served languages.

2 Overview of the PRINCIPLE project

PRINCIPLE was a two-year, EU-funded project that ran between 2019 and 2021 to identify, collect and curate high-quality LRs for the under-resourced languages of Croatian, Irish, Norwegian, and Icelandic. The Action was coordinated by Dublin City University (Ireland), and involved the Faculty of Humanities and Social Sciences of the University of Zagreb (Croatia), the University of Iceland (Iceland), the National Library of Norway (Norway), and Iconic Translation Machines Ltd. (Ireland).[6] The main focus of the project was on providing new data to improve the two Digital Service Infrastructures (DSIs) of eJustice and eProcurement, due to their strategic importance across the EU, in individual Member States and in the associated countries of Iceland and Norway. PRINCIPLE collected LRs that were general in nature as well as in a range of other

1 PRINCIPLE is an acronym for "Providing Resources in Irish, Norwegian, Croatian and Icelandic for the Purposes of Language Engineering". All the websites and online information sources mentioned in the article were last accessed on 15 September 2022.

2 In recognition of the specificities and priorities of the Norwegian partner and organisations involved in the PRINCIPLE project, data sets were collected only in Bokmål, but not in Nynorsk.

3 The catalogue of the European Language Grid (ELG) is a comprehensive repository of resources, tools, and services for all the languages of Europe (understood as a geographical rather than a political or institutional entity) and was created as part of a large-scale initiative. This collection is available here and was used as a basis for the cross-language comparison presented by the European Language Equality (ELE) Dashboard, which can be consulted here. The ELG repository and the ELE Dashboard show the comparatively disadvantaged status of Croatian, Irish, Norwegian and Icelandic in terms of technological support and availability of digital resources with respect to most of the official European languages.

4 See here.

5 See here.

6 In June 2020, Iconic Translation Machines Ltd. was acquired by the RWS Group.

domains, such as eHealth, meteorological reports, and quality control. This article focuses exclusively on eJustice and the legal domain.[7]

State-of-the-art, domain-adapted NMT engines were built by the project partner Iconic for a number of early adopters (EAs) in Croatia, Ireland, Norway, and Iceland. EAs were public bodies and private organisations that collaborated with the project by sharing their own LRs; in return for contributing digital data sets to the project, they were offered dedicated MT systems and technical support for the duration of the project. The development and subsequent evaluation of these MT systems according to the specific use cases selected by the EAs served the purpose of validating the quality and demonstrating the actual value of the LRs collected by the project. Once the quality and effectiveness of the LRs had been verified, the data sets were used to improve eTranslation and shared with the wider community (subject to applicable licensing restrictions stipulated by the data providers, as discussed in more detail in Section 4.1).

3 The multilingual dimension of the European legal framework

Europe has a unique, multilingual body of law, within a framework comprising multiple national legal systems. Croatian and Irish are official languages of the EU, while Icelandic and Norwegian are spoken in countries that are part of the European Free Trade Association (EFTA). The EU and the EFTA are the foundation of intra-European trade. The EU as we know it today developed from the European Coal and Steel Community (ECSC) and the European Economic Community (EEC), which were established in the 1950s with the objective of boosting trade and enhancing European integration between the six original founding nations. Today, the Court of Justice of the European Union (CJEU), which represents one of the use cases discussed in Section 5.3.1 in relation to Croatian, is responsible for ensuring that EU law is interpreted and applied consistently across all EU Member States and settles legal disputes between national governments and EU institutions. Established in 1960, EFTA today comprises four state parties: Iceland, Norway, Switzerland, and Liechtenstein. EFTA Member States actively cooperate with the EU on some of its core policies. All EU and EFTA Member States are parties to the Council of Europe, an international organisation made up of 46 states with the aim of fostering democracy, human rights, and a common European cultural identity.

Such a close level of cooperation across multiple legal layers surpasses the classical view of national legal systems operating as airtight silos. European states work together, often relying on common legal instruments. In the past, the law of the colonising power was imposed on foreign subjects (as was the language). The *Corpus iuris civilis*, the civil law code adopted by the Roman emperor Justinian, was originally written in Latin and enforced across a group of territories far larger than that of the current EU. This same code was subsequently translated into Greek (Way, 2016), leading to modern Balkan and Greek law, while its Latin version constituted the basis of modern law in almost all the countries of continental Western Europe, even though Latin was no longer the language spoken by the general population in those territories (Radding & Ciaralli, 2006). Interestingly, in modern times, this plurinational legal model has dramatically changed in the European context. Contemporary European legal frameworks do not rely exclusively on a common *lingua franca*, but presuppose a form of multilingual cooperation, in which various languages with equal legal status are used to communicate.

This state of affairs poses significant challenges from a legal perspective. Firstly, because the meaning associated with certain legal terms is often open to dispute, even in a monolingual context. An example of a strongly contested maxim is the Latin saying *in claris non fit interpretatio*, i.e., where the letter of the law is clear, no interpretation is needed, given that some terms can have different interpretations. Secondly, as Kahaner (2005, emphasis original) describes, legal translators should seek to provide "*literate* rather than *literal* translations" (in Wolff, 2011). In line with this well-established doctrine, which was also reflected in the use cases discussed in this paper, the translated text should normally, although not necessarily, perform

7 The Digital Service Infrastructures include a range of online platforms and technologies, implemented as a result of large-scale investments by the EU, that deliver networked cross-border services for citizens, businesses and public administrations. Specifically, eJustice refers to the use of information and communication technologies to facilitate and improve citizens' access to justice-related and legal services, eProcurement focuses on services enabling EU companies to respond to public procurement procedures from contracting entities in any member state, and eHealth supports cross-border interactions between healthcare providers on the one hand and between citizens and healthcare providers on the other.

the same function as the original text, avoiding words that have no universally agreed legal meaning. For these reasons, European institutions have a body of professional lawyer linguists who are trained in law and proficient in multiple languages. These linguists help to ensure that official documents maintain their legal value intact in translation into other languages. Against this complex background, the present article argues that digital LRs and translation technologies capable of supporting communication in the legal domain offer opportunities with great potential.

4 Collection, curation and sharing of language resources

4.1 Overview of the procedure

Each LR contributed to PRINCIPLE was validated according to ELRC-SHARE repository guidelines.[8] The quality of the validation process was ensured by the stipulation that the validators should not have previously seen the LR, which is why this step was performed by someone other than the person uploading the LR. The ELRC Data Report template used to document the process was based on the *ELRC Guidelines for Connecting Europe Facility generic services projects V2.0*.[9] The data reports covered the following areas and were uploaded to the ELRC-SHARE repository for all LRs: compliance with the ELRC scope, quick content check, metadata validation, legal validation, content validation, and a processing report. In the context of PRINCIPLE, validating digital LRs in the four under-resourced languages meant demonstrating quality improvements over the baseline MT systems by training domain-adapted NMT engines with the newly collected LRs. These LRs are therefore bound to improve the quality of eTranslation specifically as well as prove useful in other, broader applications.

PRINCIPLE collected, validated, and shared via ELRC-SHARE more than 50 separate data sets for the languages of interest to the project, nearly half of which are in the eJustice DSI and relevant to the legal domain that is the focus of this article. Although the great majority of these LRs are bilingual parallel corpora, there are also a few monolingual and multilingual corpora, as well as a small number of glossaries.[10] The project partners consistently ensured proper handling of copyright clearance and intellectual property rights issues in relation to the digital LRs with all the relevant data providers. Most of the LRs were contributed under the "CC-BY-4.0" licence, others under "public sector information" and "non-standard/other licence terms" licences, and a few remaining LRs were contributed under other miscellaneous licences based on specific requirements and conditions set by the individual contributors. Some LRs contained proprietary and/or sensitive information and were therefore provided exclusively to the Directorate General for Translation (DGT) of the European Commission to develop eTranslation, but could not be shared with the general public via ELRC-SHARE. The majority of LRs are in plain text and Translation Memory eXchange (TMX) format,[11] with others in Microsoft Excel (XLSX) format, text with tab-separated values and text in comma-separated values, which guarantees wider reusability and interoperability to benefit the largest possible number of users and applications. The following sections provide more specific information on the LRs in the legal domain collected for Croatian, Irish, Norwegian, and Icelandic.

4.2 Language resources in the legal domain for Croatian

A total of 17 distinct LRs (16 parallel corpora and one glossary) were collected and published on ELRC-SHARE for Croatian. Most of these LRs were for the Croatian–English language pair; however, of the donated translation memories (TM), one contained translations in French in addition to English, and the glossary of legal terms included German equivalents alongside English. Nine of the 17 LRs are of particular interest

8 See here.

9 See here.

10 Similar projects and repositories focus on valuable LRs for several languages. For instance, projects such as ParaCrawl and public repositories such as OPUS provide access to LRs in a wide range of languages, including those addressed by the PRINCIPLE project, which focused on collecting new LRs, especially in the poorly supported legal domain. While we cannot rule out overlap of the contents included in each of these repositories, a systematic assessment of this overlap between our own project and these other initiatives and repositories lies outside the scope of this paper.

11 A definition of the Translation Memory eXchange format (TMX) is available here.

here because they belong to the eJustice domain, for a total of 640,348 translation units (TUs) or 27,449,885 tokens. Detailed corpus sizes are presented in Table 1.

In certain instances, where optical character recognition (OCR) software was required to extract the text from PDF documents, ABBYY FineReader was used. After converting the text to plain text format (TXT), a manual check of the adequacy of the plain text format was performed on a randomly selected sample of the data. This involved scanning the text in order to identify any technical errors resulting from OCR or other conversion between formats, such as unexpected line breaks, duplicated content, page number tags, incorrect recognition of diacritics, and other language-specific characters. This was followed by a data cleaning and pre-processing step, involving the manual correction of these errors and the removal of documents with a large number of errors. Checks for overlap in documents from different providers in the same domain were also performed and duplicates were removed.

Finally, once the text had been extracted and cleaned, where this had not been carried out by the data contributors in advance, automatic segment-level alignment was performed using Vecalign (Thompson & Koehn, 2019), a tool that reports state-of-the-art sentence alignment performance.[12] Additional details on the data cleaning, types of automatic and manual checks, and data processing performed on the Croatian resources can be found in Klubička et al. (2022). The clean and aligned plain text data was then handed over to Iconic to build bespoke MT engines, having performed additional data filtering steps as part of their development pipeline (a detailed description of this process is provided in Section 5.1; it should be noted that the pipeline was also applied to the other languages as well as Croatian). For the eJustice system, one Croatian EA donated 113,676 TUs, which were filtered down by Iconic to 100,649 TUs and eventually used for development purposes.

Table 1: List of published Croatian resources in the eJustice domain

Name	Description	Size	DSI
PRINCIPLE MVEP Croatian-English-German Glossary of Legal Terms	A glossary of legal terms from the Croatian Ministry of Foreign and European Affairs	1,485 entries	eJustice
PRINCIPLE DKOM Croatian-English Parallel Corpus of legal documents	Legal documents from the Croatian State Commission for Supervision of Public Procurement Procedures	492 TUs	eJustice
PRINCIPLE MVEP Croatian-English Parallel Corpus of legal documents	Legal documents from the Croatian Ministry of Foreign and European Affairs	113,685 TUs	eJustice
PRINCIPLE MVEP Croatian-English Parallel Corpus of Court Judgements	Legal documents from the Croatian Ministry of Foreign and European Affairs	13,335 TUs	eJustice
PRINCIPLE SDURDD Croatian-English Parallel Corpus in the legal domain	Legal documents from the Croatian Central State Office for the Development of the Digital Society	261,046 TUs	eJustice
PRINCIPLE SDURDD Croatian-English Parallel Corpus of international agreements	Legal documents from the Croatian Central State Office for the Development of the Digital Society	234,500 TUs	eJustice
PRINCIPLE FFZG Croatian-English Parallel Corpus in the eJustice domain	Documents from the legal domain from the Faculty of Humanities and Social Sciences, University of Zagreb	3,731 TUs	eJustice
PRINCIPLE MVEP Croatian-English Parallel Corpus of Decisions related to the COVID-19 disease epidemic	Documents related to the COVID-19 disease epidemic from the Croatian Ministry of Foreign and European Affairs	563 TUs	eHealth, eJustice
PRINCIPLE DKOM Croatian-English Parallel Corpus of Directives of the European Parliament and of the Council	Procurement documents from the Croatian State Commission for Supervision of Public Procurement Procedures	11,511 TUs	eJustice, eProcurement
Total		640,348 TUs	

12 See here.

4.3 Language resources in the legal domain for Irish

Six LRs in the eJustice domain were collected for Irish, all in combination with English. These LRs were related to primary legislation (463,530 TUs), secondary legislation (64,385 TUs), and ancillary sources (37,729 TUs). Detailed corpus sizes are presented in Table 2. The data were provided in a wide variety of formats. Some documents were aligned and delivered in TMX, XML-based, or Excel formats, requiring little processing beyond text extraction and normalisation of encoding. Others were not aligned (i.e., Microsoft Word, HTML, or PDF format) and needed more pre-processing, involving text extraction, normalisation of encoding, sentence splitting, TU alignment, and general cleaning. This was largely performed using Python code, custom written for PRINCIPLE. TUs were aligned using the external tool Hunalign (Varga et al., 2005). An appropriate machine-readable English-Irish dictionary was not available at the time, and therefore no dictionary was used with Hunalign. Nevertheless, its performance in the absence of such a dictionary was satisfactory. Reusable pre-written code modules (i.e., libraries) were used to support normalisation and text extraction for some of the input formats; the libraries in question (openpyxl and unidecode) are publicly available on the software repository PyPI.[13] For other file types, text was extracted using external tools (libreoffice,[14] pdftotext[15]) or custom-written Python code. All data was reformatted into UTF-8 plain text files.

Next, regardless of their source, the output text files were run through an automated error checker, which was also custom written for PRINCIPLE. This tool searched for a variety of anomalies that might indicate bad or partial alignments. Specifically, it searched for pairs containing the following errors:

- the two sides were of greatly differing lengths;
- one side contained a numeral absent from the other;
- one side was in uppercase and the other was not (because the legal text in question was observed to be case-sensitive, this check was useful in detecting misalignments, but complicated by the need to account for the fact that Irish has different rules from English for admitting lowercase letters in otherwise uppercase text);
- one or both sides consisted exclusively of non-alphabetic characters;
- one side consisted of two or more sentences where the other had only one (as the two languages are structurally different, this could be valid, but was very often a sign of misalignment);
- both sides were identical, which was often a sign of an untranslated segment.

Such errors could have been introduced by the earlier automated processing steps, or there could have been a problem with the data itself, such as typographical issues or erroneous artefacts from translation or publishing. Errors detected by the tool were investigated and corrected manually.

There was a certain amount of overlap between this process and the procedure described in Section 5.1. The process in Section 5.1 was entirely automated, and performed general purpose operations, such as tokenisation and truecasing, aimed at creating suitable input specifically for MT systems. The procedure here, however, was both automated and manual, and more concerned with general cleaning. This phase of cleaning was more tailored to the data and language pair involved and, being closer to the original data, it had the benefit of being able to inspect the raw documents if something was in question.

Lastly, the output was spot-checked manually for quality, before being handed to Iconic to build their MT systems. One Irish EA donated 387,480 TUs, which were filtered down by Iconic to 353,485 TUs that were eventually used for development purposes.

13 See here and here.

14 See here.

15 See here.

Table 2: List of published Irish resources in the eJustice domain

Name	Description	Size	DSI
PRINCIPLE Dept of Justice parallel English-Irish secondary legislation	Legal corpus of statutory instruments, including rules of court, from the Department of Justice in Ireland	35,898 TUs	eJustice
PRINCIPLE English-Irish parallel primary legislation 1960 to 1989	Corpus of primary legislation by Rannóg an Aistriúcháin	177,797 TUs	eJustice
PRINCIPLE English-Irish parallel primary legislation 1990 to 2019	Corpus of primary legislation by Rannóg an Aistriúcháin	285,733 TUs	eJustice
PRINCIPLE English-Irish parallel secondary legislation	Corpus of secondary legislation by Rannóg an Aistriúcháin	28,487 TUs	eJustice
PRINCIPLE English-Irish Houses of the Oireachtas ancillary material dataset	Corpus of data relating to the running of the Houses of the Oireachtas by Rannóg an Aistriúcháin	35,614 TUs	eJustice
PRINCIPLE English-Irish glossary of terms relating to primary legislation in Ireland	Glossary of terminology relating to primary legislation in Ireland from Rannóg an Aistriúcháin	2,115 Terms	eJustice
Total		563,529 TUs, 2,115 terms	

4.4 Language resources in the legal domain for Norwegian

A total of 1,156,999 TUs in the eJustice DSI were collected for Norwegian, along with a glossary of 42,141 English and French–Norwegian terms related to EU legislation. A breakdown of the published corpora is presented in Table 3. Of the LRs published on ELRC-SHARE for Norwegian, 38% were for the eJustice domain, and all of them were made available under the "CC0 1.0 Public Domain Dedication" licence. By far the most important data contribution (1,147,000 TUs) came from the Norwegian Ministry of Foreign Affairs (MFA) and consisted mainly of translations from English into Norwegian of EU legal documents. The Norwegian MFA also contributed the above-mentioned multilingual glossary of legal terms. A second minor contribution (67,000 TUs) came from the Norwegian Maritime Authority and consisted mainly of translations from Norwegian into English of legal and statutory documents. The translation department of the Norwegian MFA uses computer-assisted translation (CAT) tools and was therefore able to share translation memories in TMX format. The Norwegian Maritime Authority does not use CAT tools, and the National Library of Norway received a manually-aligned TMX file that was corrected and validated before the resource was published in ELRC-SHARE. The clean and aligned plain text data was handed over to Iconic, who used it to build bespoke MT engines.

Table 3: List of published Norwegian resources in the eJustice domain

Name	Description	Size	DSI
Translation memory from the Norwegian Maritime Authority	Norwegian (mainly Bokmål)–English legal and statutory documents	68,296 TUs	eJustice
Translation memories from the Norwegian MFA – 2020	English–Norwegian (mainly Bokmål) EU legal documents	1,088,703 TUs	eJustice
EU term base from the Norwegian MFA	English and French–Norwegian (mainly Bokmål) glossary relating to EU legislation	42,141 terms	eJustice
Total		1,156,999 TUs, 42,141 terms	

4.5 Language resources in the legal domain for Icelandic

Two separate LRs in the eJustice DSI were collected for Icelandic, with a total of 1,102,025 TUs. Detailed corpus sizes are presented in Table 4. Documents were manually collected from a number of relevant websites,

a task that could not be accurately automated for Icelandic using Terminotix's AlignFactory.[16] It is possible that ILSP Focused Crawler or Bitextor might achieve better results. The most common processing step was a mix of manual and automatic text extraction from PDF, HTML, and DOC(X) files and conversion to plain text format (TXT). In certain cases, the use of OCR software was required to extract the text from PDF documents; this was carried out using ABBYY PDF Transformer+. Once converted to plain text format, a manual check, data cleaning, and pre-processing were performed. Any technical errors due to OCR or other conversion between formats were corrected. AlignFactory was used to automatically align sentences and create a TM in TMX format. The TUs were subsequently randomised. Eventually, Iconic used 1,095,312 TUs to build the eJustice NMT engine.

Table 4: List of published Icelandic resources in the eJustice domain

Name	Description	Size	DSI
The Translation Centre of the Icelandic Ministry for Foreign Affairs – Gullsarpur	English–Icelandic translation memory from the Translation Centre of the Icelandic MFA, containing all domains	1,065,736 TUs	eJustice
Government Offices of Iceland – Legislation and regulations	English–Icelandic translation memory created from documents on the Icelandic and English websites of the Government Offices of Iceland.	36,289 TUs	eJustice
Total		1,102,025 TUs	

5 Development of bespoke neural machine translation engines in the legal domain

5.1 Architecture of the machine translation engines

A number of MT engines were built by Iconic, including baseline systems trained on existing ELRC-SHARE data together with bespoke, domain-adapted EA neural engines developed using the domain-specific data collected by the project, as described in Section 4. A baseline MT system serves as a good quality, defined starting point to compare subsequent retrained and improved versions of the system, and is thus useful to verify any increase in translation quality over time. The NMT systems are 11x3 Transformer base models (Vaswani et al., 2017). The tokens of the training, evaluation, and validation sets were segmented into subword units using BytePair Encoding (BPE) (Sennrich et al., 2016). The recommended setup described in Vaswani et al. (2017) was followed, and the early stopping criteria is based on cross entropy. Iconic's extensive data cleaning and preparation pipeline, described in full in Gupta et al. (2019), ensured that any noisy training data was eliminated and only reliable data was used to train the MT engines. In adherence with widely accepted standards in the MT industry, the pipeline in question included the following processes:

- character and encoding cleaning;
- punctuation and digit filtering;
- copy filtering;
- duplicate removal;
- a length-based filter;
- a language-based filter;
- do-not-translate word replacement;

16 See here.

- a processing pipeline, including tokenisation,[17] truecasing,[18] and all the standard operations needed to ensure that the statistical calculations underpinning the MT systems are not unduly affected by incorrect counting of words.

Once thoroughly cleaned, the resulting data was processed with Iconic's proprietary technology and methodology to build a number of baseline systems and EA NMT engines. In the course of developing the first batch of systems, the cleaning pipeline was adjusted to adapt to the specific formatting used by resource contributors and ensure that as much data as possible was used for training the engines and not erroneously discarded. Additionally, based on feedback received from the EAs, the pipelines of newly trained engines were reviewed and components to tackle language-specific issues (e.g., incorrect capitalisation) were improved, added or removed to fit the needs of each EA and increase the overall quality of the MT engines.

5.2 Data used for training the baseline machine translation systems

In total, seven baseline systems were built for the following language combinations of interest to the EAs in the four countries covered by the project, using publicly available data taken from ELRC-SHARE after the PRINCIPLE partners had reviewed and checked it for relevance, using the general procedures and techniques described in Gupta *et al*. (2019): bidirectional English↔Croatian (EN↔HR), monodirectional English→Irish (EN→GA), bidirectional English↔Norwegian (EN↔NO), and bidirectional English↔Icelandic (EN↔IS). From a total of 3,892,777 TUs collected, 3,337,608 were selected to train the English↔Croatian baseline engines (the list of filters and method of selecting the TUs is described in Section 5.1). Of the 901,421 TUs available for English→Irish, 588,663 were eventually used to train the baseline engine. To train the English↔Norwegian baseline systems, 1,140,351 TUs were used from the original 1,964,961 that were potentially available. Finally, for English↔Icelandic, the original number of 801,283 TUs made available was reduced to 702,139.

5.3 The importance of building custom neural machine translation engines in the legal domain for early adopters

In total, twelve bespoke, state-of-the-art NMT engines in a number of domains were trained for various EAs during the term of the PRINCIPLE project in the following language combinations: two English→Norwegian, one Croatian→English, two English→Icelandic, two Icelandic→English, three English→Irish, and two English→Croatian. These bespoke, domain-adapted NMT engines offered to the EAs for the duration of the project, with Iconic's full technical support, acted as a valuable incentive to encourage organisations in Croatia, Ireland, Norway, and Iceland to contribute their LRs. A number of studies explain the challenges and benefits of domain adaptation of MT systems to successfully deal with specialised texts (e.g., Chen et al., 2017; Chu et al., 2017; Chu et al., 2018; Chu & Wang, 2018; Etchegoyhen et al., 2018).

In the language and translation service industry, it is widely recognised that generic, one-size-fits-all MT engines such as those available free of charge on the Internet are generally not appropriate for use for professional purposes due to poor output quality, limited flexibility, and security and data confidentiality concerns. For its work on the PRINCIPLE project, Iconic tailored its MT solutions to the specific requirements of the EAs, based on the needs of the under-resourced languages, the individual content types, deployment scenarios, and envisaged user interactions. This work encouraged the uptake and further adoption of translation technologies.

This, in turn, fostered the engagement of EAs in the long term for the sustainable detection, collection, processing, and sharing of digital LRs. This longer-term engagement, beyond the lifetime of the PRINCIPLE project, is essential to offset the imbalance in terms of LR availability between low-resourced and better

17 In natural language processing, tokenisation refers to the process of separating strings of characters to identify actual words or their meaningful parts, as a prerequisite for the subsequent translation. For example, tokenisation identifies "I" and "'m" as two distinct (parts of) words that contribute to the overall meaning of the English phrase "I'm", which despite consisting of continuous characters in spelling, is correctly broken down into two words by the system and properly analysed.

18 Truecasing involves using the correct capitalisation of words, as required by the spelling rules of the language being processed. This involves, for instance, ensuring that (uppercase) "The" and (lowercase) "the" are correctly interpreted and processed by the system as essentially the same word, i.e., the definite article in English, despite the variation in spelling due to their different positions at the beginning of and within a sentence, respectively.

supported languages. This is particularly important at a time when a wide range of professionals, academics, researchers, and Internet users in general increasingly use MT in a variety of fields, often with no clear understanding of the potential and, more importantly, of the limitations of current translation technologies (e.g., Bowker & Ciro, 2019; O'Brien & Ehrensberger-Dow, 2020).

5.3.1 The Ministry of Foreign and European Affairs of the Republic of Croatia, English to Croatian

The Ministry of Foreign and European Affairs of the Republic of Croatia required the development of a bespoke MT system in the legal domain for English to Croatian, for the use case of judgements of the CJEU. To train this system, Iconic made use of data coming from the eJustice domain, enriched with additional LRs from other domains. In this instance, it was not possible to develop a bespoke MT system from scratch due to an insufficient volume of Ministry of Foreign and European Affairs of the Republic of Croatia data. A client-tailored MT engine was therefore built by fine-tuning the baseline EN-HR engine, using data provided by the Ministry of Foreign and European Affairs of the Republic of Croatia.[19]

5.3.2 The Rannóg an Aistriúcháin and the Department of Justice of Ireland, English to Irish

The Rannóg an Aistriúcháin's focus is almost entirely on legal translation, i.e., primary and secondary legislation, together with a small amount of more general purpose work (press releases, annual reports, etc.). The approximately 20 translators working at the Rannóg an Aistriúcháin only began using CAT tools in recent years, so the bulk of their data comes in custom or unaligned legacy formats. Legal data sets were also collected from the Department of Justice of Ireland, where a single member of staff handles the in-house translation needs, and whatever material they are unable to work on due to time constraints is outsourced.

The use cases of the Rannóg an Aistriúcháin and the Department of Justice of Ireland required the development of a domain-adapted, bespoke MT system for the domain of eJustice for English to Irish. Due to the amount of data collected, Iconic proposed to produce one MT engine for both these EAs, for which 1,028,844 TUs in total were used, less than a third of which came from the eJustice domain. In this instance, a robust, client-tailored MT system was built by fine-tuning the English–Irish baseline engine, using data provided by the Rannóg an Aistriúcháin, the Department of Justice of Ireland, and Ireland's Department of Culture, Heritage and the Gaeltacht (DCHG).

5.3.3 The Norwegian Ministry of Foreign Affairs, English to Norwegian

The use case of the Norwegian MFA required a domain-adapted bespoke MT engine for the eJustice domain from English to Norwegian. For the development of a domain-tailored engine, the Norwegian MFA provided 1.1 million TUs of legal documents translated from English to Norwegian in TMX format. This was deemed a sufficient volume of training material with which to build a robust MT engine.

5.3.4 The Translation Centre of Iceland's Foreign Ministry, English to Icelandic

The Translation Centre of Iceland's Foreign Ministry had two MT engines built for the legal domain, to translate between English and Icelandic in both directions. Eventually, the EN→IS engine was fully evaluated, including with human evaluation based on post-editing, while the IS→EN engine was not tested beyond automatic evaluation. As a result, the discussion in Section 6.6.4 is limited to the EN→IS engine, for which the more comprehensive evaluation was performed. It should be noted that the Translation Centre of Iceland's Foreign Ministry requested that no additional training data be integrated with their own for the development of their MT engine, possibly due to the sufficient volume of training data they provided.

19 It is not always necessary, or possible, to build a brand new MT engine from scratch; indeed, in many cases there will be insufficient amounts of suitable data to justify this process. In such circumstances, an alternative is to fine-tune an existing system, for example by optimising the internal parameters and weights to obtain noticeable quality improvements over the baseline.

6 Evaluation of the machine translation engines and validation of the language resources

6.1 Approach to machine translation evaluation in PRINCIPLE

This section gives an overview of the overall approach, general methodologies, and materials (including test sets and guidelines with evaluator instructions) specifically designed and used to evaluate the PRINCIPLE MT engines. Castilho et al. (2018) review the strengths and weaknesses of a range of human and automatic MT evaluation methods, suggesting that, when possible, a combination of human/manual and automatic approaches is likely to offer advantages, in the interests of a realistic balance between speed, cost, and overall reliability of the evaluation.

When designing the MT evaluation tasks and materials for the project, special attention was paid to combining automatic evaluation metrics (AEMs) with suitable human/manual methods, so as to maximise their respective advantages. A key priority in this respect was to match the needs and intended use cases of the EAs, so that the evaluation setup adopted for each of them could provide meaningful results to improve the MT engines. In particular, we wanted to guarantee flexibility, to make sure that the chosen evaluation protocols corresponded to the specific use cases of the individual EAs, as well as to the time and resources each EA was able to invest in the human element of the evaluation.

6.2 State-of-the-art, automatic evaluation metrics used

AEMs are widely used, mainly because they are easy, quick and cheap, especially in comparison to the much slower and more expensive human approaches (Way, 2018). For the automatic evaluation, the decision was made to use a suite of well-established metrics, namely BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005; Lavie & Agarwal, 2007), Translation Edit Rate (TER) (Snover et al., 2006), and chrF (Popović, 2015).

The string-based metric BLEU has been a *de facto* standard in MT evaluation, both in industry and academia, for several years now, primarily because it is fast and intuitive. While some researchers have expressed reservations about its actual value and usefulness (e.g., Callison-Burch et al., 2006; Denkowski & Lavie, 2012), its long-lasting impact on the MT community cannot be denied. Compared to BLEU, METEOR provides a more fine-grained and delicate evaluation at the lexical level, thus reducing bias due to the oracle reference translations. Adopting a different approach, TER is an edit metric based on the number of substitutions, insertions, deletions, and shifts of blocks of contiguous words required to modify the MT output in such a way that it completely matches the reference, and is computed as the ratio of this number of operations to the length of the sentence. Hence, the lower its score, the better the results, as a lower score indicates fewer modifications were needed to match the reference translation. The rationale behind this metric is quite simple for non-experts in MT to understand, as it provides an approximation of the amount of post-editing needed by an end-user. Finally, chrF operates at the subword level, by comparing the character *n*-grams (i.e., a string of *n* adjacent letters that are part of a word) in the hypothesis with those in the reference, thereby affording more delicacy at the morphogrammatical level for highly inflected languages. This method has been reported to correlate very well with human direct assessment scores based on adequacy and partly on fluency (e.g., Bojar et al., 2017).

For some of the automatic evaluations, the BLEU scores were obtained using the SacreBLEU toolkit (Post, 2018)[20] to ensure standardised reporting of results, and are therefore comparable and reproducible. In this article, the notation SacreBLEU is used to refer to BLEU scores computed using this toolkit.The other automatic evaluations based on BLEU, METEOR and TER (Tables 11, 17, 19, 24 and 27) were performed using the MultEval package, which is a user-friendly implementation of Clark et al. (2011).[21] MultEval produced the same scores obtained by running the metrics as stand-alone programs, and no tokenisation was performed. MultEval called METEOR and TER using library Application Programming Interface calls, and a regression test was calculated against gold-standard scores produced by these metrics in stand-alone mode.

20 See here.

21 See here.

6.3 Automatic evaluation of baseline machine translation systems

Baseline systems were created with LRs already available on ELRC-SHARE for the language combinations of interest, without using any of the data collected by PRINCIPLE. The baseline evaluations were performed on test sets of 3,000 TUs. These 3,000 TUs were extracted from the original training data available on ELRC-SHARE and therefore were not used for training the engines. Such test sets were a randomised representation of the training data, and therefore suitable for evaluating generic MT baseline engines. The translations of the respective test sets were retrieved using commercial MT systems (Google Translate[22] and Microsoft Bing[23]) in November 2020, except for the second English into Croatian engine, for which translations were retrieved in May 2021.

Table 5: Automatic evaluation of the initial EN→HR baseline MT engine

Engine	SacreBLEU	TER	METEOR	chrF
Initial EN→HR baseline	42.4%	49.8%	49.3	64.2%
Google Translate	35.0%	55.6%	42.9	59.2%
Microsoft Bing	34.5%	54.8%	42.6	58.2%

The initial EN→HR baseline engine was retrained for improved performance, and subsequently evaluated on a different blind test set of 2,000 segments, with the results shown in Table 6.[24]

Table 6: Automatic evaluation of the retrained EN→HR baseline MT engine

Engine	SacreBLEU	TER	METEOR	chrF
Retrained EN→HR baseline	47.8%	42.0%	55.2	67.7%
Google Translate	34.7%	53.9%	43.2	59.1%
Microsoft Bing	35.6%	51.2%	44.6	59.8%

Table 7: Automatic evaluation of the EN→GA baseline MT engine

Engine	SacreBLEU	TER	METEOR	chrF
EN→GA baseline	58.4%	36.3%	64.9	75.3%
Google Translate	39.0%	52.3%	47.9	64.2%
Microsoft Bing	40.2%	53.0%	50.1	63.1%

Table 8: Automatic evaluation of the EN→NO baseline MT engine

Engine	SacreBLEU	TER	METEOR	chrF
EN→NO baseline	47.1%	46.9%	54.0	66.8%
Google Translate	35.2%	57.9%	43.0	58.0%
Microsoft Bing	31.8%	59.2%	40.7	55.4%

Table 9: Automatic evaluation of the EN→IS baseline MT engine

Engine	SacreBLEU	TER	METEOR	chrF
EN→IS baseline	45.4%	43.1%	55.7	66.3%
Google Translate	31.0%	54.2%	42.4	56.9%
Microsoft Bing	32.6%	51.7%	43.7	56.5%

22 See here.

23 See here.

24 We acknowledge that this new test set renders the evaluation of the second baseline not directly comparable with the evaluation of the first baseline. This choice had to be made because, in the time between the first baseline and the second, a technical issue prevented Iconic from being able to access the first engine and test set originally used.

6.4 Preliminary automatic evaluation of the bespoke early adopter machine translation engines

6.4.1 EN→HR engine for the Ministry of Foreign and European Affairs of the Republic of Croatia in the legal domain

A test set of 1,500 segments was withheld from the training data received from the Ministry of Foreign and European Affairs of the Republic of Croatia, and the evaluation scores obtained from this test set are shown in Table 10.[25]

Table 10: Initial evaluation of the Ministry of Foreign and European Affairs of the Republic of Croatia eJustice EN→HR MT engine

Engine	SacreBLEU	TER	METEOR	chrF
Iconic Ministry of Foreign and European Affairs of the Republic of Croatia Engine	51.1%	39.0%	58.6	69.9%
Google Translate	37.9%	51.6%	46.4	60.8%
Microsoft Bing	31.7%	57.4%	41.2	55.1%

Since, due to time constraints, it was not possible to perform human evaluation (Section 6.6) of this EA MT engine, an additional automatic evaluation was performed on a different test set of 500 sentences, for a total of 23,086 tokens, also including eTranslation. As Table 11 shows, the Iconic engine clearly outperformed Google Translate, Microsoft Bing, and eTranslation.

Table 11: Additional evaluation of the Ministry of Foreign and European Affairs of the Republic of Croatia eJustice EN→HR MT engine

Engine	BLEU	METEOR	TER	chrF
Iconic Ministry of Foreign and European Affairs of the Republic of Croatia Engine	28.0%	25.8	53.0%	56.4%
Google Translate	21.4%	22.2	61.7%	52.1%
Microsoft Bing	12.7%	16.4	71.6%	22.7%
eTranslation	25.8%	24.3	56.8%	55.5%

6.4.2 EN→GA engine for the Rannóg an Aistriúcháin in the legal domain

A test set of 2,000 segments was withheld from the training data received from the Rannóg an Aistriúcháin to compute the AEM scores shown in Table 12.[26]

Table 12: Results of the Rannóg an Aistriúcháin eJustice EN→GA MT engine on the Rannóg an Aistriúcháin test set

Engine	SacreBLEU	TER	METEOR	chrF
Iconic EA Engine (Rannóg an Aistriúcháin test set)	73.6%	19.8%	77.1	81.7%
Google Translate	48.3%	42.9%	54.9	64.4%
Microsoft Bing	55.7%	34.4%	61.6	69.3%

6.4.3 EN→GA engine for the Department of Justice of Ireland in the legal domain

A test set of 1,000 segments withheld from the training data received from the Department of Justice of Ireland was used to further test the MT system. The results of this evaluation are shown in Table 13.[27]

25 Translations retrieved in July 2021.

26 Translations retrieved in July 2021.

27 Translations retrieved in July 2021.

Table 13: Results of the Rannóg an Aistriúcháin eJustice EN→GA MT engine on the Department of Justice of Ireland test set

Engine	SacreBLEU	TER	METEOR	chrF
Iconic EA Engine (Rannóg an Aistriúcháin Engine, Department of Justice of Ireland Test set)	50.6%	40.5%	59.1	70.3%
Google Translate	47.6%	45.8%	53.3	65.7%
Microsoft Bing	49.2%	42.2%	57.5	67.4%

6.4.4 EN→NO engine for the Norwegian Ministry of Foreign Affairs in the legal domain

Due to overlap between the EA training data received from the Norwegian MFA and data previously collected from ELRC-SHARE for the development of a baseline system, cross evaluation for the newly built Norwegian MFA MT engine was calculated on a cleaned test set of a reduced volume of TUs. Approximately 733,000 TUs of EA data received from the Norwegian MFA had previously been uploaded to ELRC-SHARE and used in PRINCIPLE to train a baseline engine for Norwegian, the evaluation results of which are shown in the second row of Table 14. Duplications between baseline training data and the 2,000 TU test set withheld from the Norwegian MFA data were removed in order to conduct a fair and unbiased cross comparison of engines. The deletion of duplicates resulted in an internal Norwegian MFA test set of 897 TUs.[28]

Table 14: Cross evaluation of the Norwegian Ministry of Foreign Affairs EN→NO bespoke engine, baseline EN→NO engine, Google Translate, and Microsoft Bing

Engine	SacreBLEU	TER	METEOR	chrF
Norwegian MFA engine	66.5%	28.7%	71.9	81.0%
Norwegian baseline	54.3%	38.7%	61.3	72.5%
Google Translate	31.3%	56.3%	42.3	55.2%
Microsoft Bing	43.7%	48.0%	51.2	64.0%

6.4.5 EN→IS engine for the Translation Centre of Iceland's Foreign Ministry in the legal domain

Table 15 shows the evaluation scores obtained from a test set of 2,000 segments withheld from the training data received from the Translation Centre of Iceland's Foreign Ministry.[29]

Table 15: Cross evaluation of the Translation Centre of Iceland's Foreign Ministry EN→IS bespoke engine, baseline EN→IS engine, Google Translate, and Microsoft Bing

	SacreBLEU	TER	METEOR
Translation Centre of Iceland's Foreign Ministry engine	60.1%	35.9%	64.4
Icelandic baseline	57.8%	38.2%	37.8
Google Translate	34.7%	57.4%	41.2
Microsoft Bing	23.2%	68.6%	30.9

6.5 Ad-hoc test sets

All the EAs received specific guidelines for the preparation of test sets for the follow-up automatic and human MT evaluations. For this purpose, 500 additional segment pairs not already used to train the MT systems were set aside. We requested that the test sets be representative of the texts that the EAs eventually intended to translate with the MT engines built and deployed for them by Iconic, with regard to the use of specialised terminology, sentence length, structure, and complexity, for example. In addition, we stipulated that the human reference translation in the target language should not be obtained through MT post-editing. In particular, we were mindful of Zhang and Toral's (2019) investigation into the impact of *translationese* on test data for MT

28 Translations retrieved in November 2020.

29 Translations retrieved in November 2020.

evaluation. This study analysed test sets from the news shared task from the Workshop on Machine Translation (WMT), concluding that the presence of translationese in the test sets inflated the human evaluation scores for MT systems, occasionally affecting system rankings.

6.6 Human evaluation and additional follow-up automatic evaluation of the bespoke early adopter machine translation engines

Since the aim of the human evaluation supplementing the automatic metrics was to be as flexible and adaptable to the use cases and specific situations of the EAs as possible, a range of options for human evaluation were proposed to the EAs, namely (i) comparative ranking, (ii) adequacy and fluency assessment, (iii) direct assessment, (iv) comprehension test, (v) post-editing, and (vi) MT error analysis (for detailed explanations of these types of human evaluation, see Castilho *et al.*, 2018). The idea behind this approach was to offer the EAs a menu-based set of options, from which they could pick and mix the type(s) of human evaluation that were of greatest interest and value to them, in terms of how to assess the quality and usefulness of the MT systems built for them as part of the project. The EAs were able to choose either one or a combination of the available options. For each of the six human evaluation options, the guidelines provided to the EAs gave specific information indicating, inter alia, the level of difficulty or complexity, an estimate of the person-time and amount of commitment required of the evaluators, pros and cons, ideal application scenario, and expected usefulness. After sharing this set of options with the EAs, calls were made to the EAs to explain and discuss the options in more detail and agree on the specific evaluation protocol to be adopted on an individual basis. These follow-up sessions helped answer the EAs' questions and queries, and especially for ensuring that the methods eventually chosen were relevant to the specific use cases. It was also important to discuss the type and level of commitment required of the evaluators to ensure that the evaluation was feasible and could be conducted within the time available for the preparation of this report.

The EAs involved in this part of the study had limited or no prior experience with human MT evaluation, and therefore appreciated the coaching they received on the finer experimental points of conducting human MT evaluation. This facilitated the final agreement on a convenient and effective evaluation setup that was tailored to their specific circumstances and constraints. Before running the actual evaluations, pilot tests were conducted to check the clarity of the instructions and the materials to be used. All the evaluators provided by the EAs were experienced linguists and translators (some were in-house staff, others long-term freelancers with well-established collaborations) and received a number of specific requests in relation to the evaluation. These involved, for example, always considering the professional translation quality standard normally expected in their organisations and/or in the translation projects that they typically worked on. In cases when more than one judge took part in the evaluations, the teams were asked to perform their evaluation tasks in isolation, i.e., without communicating with colleagues or discussing their opinions with others while the evaluation was in progress. Finally, for all the evaluation types, the evaluators were encouraged to add comments and explain their opinions with regard to the evaluation, for the purpose of gathering additional qualitative evaluation data. Croatian is not included in this section, as the Croatian EAs did not perform any evaluation on NMT engines for the legal domain.

6.6.1 Evaluation for the Rannóg an Aistriúcháin, EN→GA

Two different test sets were used to perform the detailed evaluation of the Rannóg an Aistriúcháin MT engine. The first test set (1) included the source and the human reference and was used to compute the automatic metrics only. The second test set (2) consisted of source sentences only and was used to run the human evaluation task. Table 16 shows the statistics for each of these two test sets.

Table 16: Statistics for test set 1 and test set 2 for Rannóg an Aistriúcháin, EN→GA

Test set	Sentences	Tokens
1 - Source (AEMs)	500	13,622
2 - Source (human)	407	10,581

The reason for the two test sets was that the EA translator who translated test set 1 (providing the human reference for test set 1) was the same translator who conducted the human evaluation. To avoid any potential

bias (e.g., the translator remembering translation choices for particularly challenging phrases), we decided to use a second test set (2), one that had not been translated previously by the Rannóg an Aistriúcháin translator, and compared the translation times for post-editing TM matches against MT output.

6.6.1.1 Follow-up automatic evaluation

The automatic metrics were calculated for test set 1, with the human translation as reference. From the BLEU, METEOR and TER scores in Table 17, we can observe that Iconic's engine largely outperformed both Google Translate and Microsoft Bing when translating from English into Irish. The Iconic engine also outperformed eTranslation, although by a smaller margin, according to BLEU and METEOR (but not TER).

Table 17: Automatic metrics comparison for Rannóg an Aistriúcháin's test set 1, EN→GA

Test set 1	BLEU	METEOR	TER
Iconic EA Engine	73.5%	51.4	19.8%
Google Translate	47%	36.3	38.5%
Microsoft Bing	54.6%	40.4	31.8%
eTranslation	72.9%	50.3	18.9%

6.6.1.2 Human evaluation based on post-editing

For the human evaluation, the Rannóg an Aistriúcháin decided to compare the post-editing of the MT output against their normal workflow where TMs are used. Test set 2 was then translated with Iconic's bespoke NMT engine and by the Rannóg an Aistriúcháin's TM with matches over 60%, a threshold agreed to by the EA and the PRINCIPLE partner in charge of running the human evaluation. Based on the managers' preference for their own internal translation service, this threshold was determined as being more reflective of the material with which they typically work and for which they would consider introducing MT. The MT output and TM matches were mixed in four different files and sent to two translators.[30] The translators then uploaded the source (English) and target (Irish) files to their CAT tool, Trados. The translators were told to:

- Disable TM matches in Trados. (The researchers needed to know whether post-editing MT output or modifying TM matches is faster.)
- Keep track of the time taken to translate each file. (The translators chose to do this manually with a stopwatch because a Trados plugin would have been problematic to install.)

The results of the productivity task in terms of time are given in Table 18. After the task was completed, a post-task questionnaire was sent to the translators to collect their comments, and some of their most interesting responses are discussed below.

Table 18: Time spent on MT post-editing and translation with TM matches, EN→GA

Time	MT post-editing		TM	
(hh:mm:ss)	T1	T2	T1	T2
File 1	04:15:00	01:40:56	03:00:00	01:57:30
File 2	03:30:00	00:48:13	03:30:00	01:24:27
Total	07:45:00	02:29:09	06:30:00	03:21:57

Using MT was faster than using the TM matches for translator 2 (2:29 hours vs. 3:21 hours, respectively), whereas translator 1 spent more time post-editing MT than using the TM matches (7:45 hours vs. 6:30 hours, respectively). One section of this human evaluation asked the translators for their impressions; some of the most interesting answers are included below.

30 Note that the outputs were mixed when sent to the translators: while translator 1 saw source sentences 1–100 with the MT output, translator 2 saw source sentences 1–100 with the TM matches.

Were you able to identify which file was TM and which was MT?

T1: Yes, because there were meaningless words in the target text which could only have come from machine translation.

T2: Yes, the TMs retained the specific legislative terminology in use in the Rannóg an Aistriúcháin, which can differ from other forms of pragmatic translation (as in use in annual reports, for example).

Would you use MT for post-editing again at the Rannóg an Aistriúcháin?

T1: I would use MT again and I think it would assist the translator in his/her work.

T2: I likely would. For shorter sentences it was very good. For longer sentences it had errors but it is a help to have certain clauses of the sentence correct. It is easier to move clauses around and correct terms and grammar rather than starting from scratch.

These results indicate that the Rannóg an Aistriúcháin translators, who had never used MT professionally before, think that they would benefit from using MT output, in the interest of increasing productivity, especially with shorter sentences and as they improve their post-editing skills. The translators reported never having received formal training in post-editing prior to their participation in this evaluation.

6.6.2 Evaluation for the Department of Justice of Ireland, EN→GA

The evaluation for the Department of Justice of Ireland was based on a test set of 248 sentences, corresponding to 2,703 tokens on the source side.

6.6.2.1 Follow-up automatic evaluation

The AEM scores in Table 19 were calculated using the human translation as reference. Again, from the BLEU, METEOR and TER scores we can observe that Iconic's engine outperformed both Google Translate and Microsoft Bing. Interestingly, for this specific test set, eTranslation outperformed Iconic's engine across all AEMs. As the test set was not part of the training data, one possible explanation is that it was very different from the type of content provided to train the engine. Moreover, given that we have access to neither the list of resources used to train the eTranslation engine nor the specifics of the engine, it is difficult to determine the origin of the differences in scoring. However, as reported in Section 6.6.2.2, the human evaluation of the system for adequacy and usefulness shows that the Iconic engine was fit for purpose.

Table 19: Automatic metrics comparison for the Department of Justice of Ireland's test set, EN→GA

Engine	BLEU	METEOR	TER
Iconic EA Engine	45.0%	35.1	40.4%
Google Translate	38.9%	31.5	45.8%
Microsoft Bing	41.8%	33.5	41.8%
eTranslation	49.9%	36.9	38.5%

6.6.2.2 Human evaluation based on error markup, adequacy, and usefulness of machine translation output

The human evaluation conducted with the Department of Justice of Ireland consisted of an error markup task, adequacy scoring, and a question to ascertain whether the translator thought the MT output would be useful to the translation workflow. For the error-based evaluation, the EA chose a taxonomy with five error types from the Multidimensional Quality Metrics Taxonomy (Lommel *et al.*, 2014) that best suited their needs; the error types selected were mistranslation, omission, grammar, register, and terminology. For each sentence, the translator could tag any number of errors and any issue type, which means that more than one issue type could be tagged per sentence, and the same issue type could be tagged more than once. The translator was asked to select "no issues" where the sentence contained no errors. Table 20 shows the results for the error markup. We note that this EA was only able to provide one translator for the human evaluation in the relevant time frame. However, we believe that this evaluation is still relevant, especially because the translator in question was the only in-house professional translator working at this particular EA. In other words, for the purposes

of the project, we were fortunate enough to be able to work with (and rely on the evaluation of) the only in-house professional translator working at this particular institutional early adopter.

In total, 550 errors were tagged in the entire MT output that was examined. The most common error was *mistranslation* (41% of all errors), followed by *terminology* (27% of all errors). The error types, *omission* and *grammar*, follow with a similar count and percentage, while *register* has very few cases, which is interesting for legal texts that typically employ a rather specific and peculiar style. Finally, 28 of the 248 sentences did not contain any errors, and they could be used directly as provided by the MT engine.

Table 20: Error annotation results for the Department of Justice of Ireland's test set, EN→GA (the percentage of errors is the total of error types divided by the total number of errors, i.e., 550)

Errors	No issues	Mistranslation	Omission	Grammar	Register	Terminology
Count	28	227	79	93	2	149
% (Error)	-	41.27	14.36	16.91	0.36	27.09

After tagging errors, the translator assigned an adequacy score to each output sentence on a 1–4 Likert scale based on the question, "*How much of the meaning of the source text is in the machine translation output?*". Table 21 shows the results for the adequacy judgement.

Table 21: Adequacy results for the Department of Justice of Ireland's test set, EN→GA

Adequacy	None (1)	Little (2)	Most (3)	All (4)
Count	3	23	130	91
%	1.21	9.27	52.42	36.69

The results for adequacy shown in Table 21 indicate that the translator assessed over 89% of the MT output sentences as containing most or all of the meaning of the respective source segments. Finally, the translator was also asked whether the MT output was considered useful or not, with the question, "*Is the machine translation output a serviceable preliminary translation for post-editing?*". Table 22 shows the answers to this question, which indicate that the translator believed that over 83% of the MT output sentences were useful as a basis for subsequent post-editing; this result is consistent with the previous finding for adequacy.

Table 22: Usefulness results for the Department of Justice of Ireland's test set, EN→GA (the percentage is the total number of "Yes" and "No" answers divided by the number of sentences)

Serviceable	Yes	No
Count	208	40
%	83.87	16.13

6.6.3 Evaluation for the Norwegian Ministry of Foreign Affairs, EN→NO

This section presents the results of the in-depth evaluation of the custom NMT engine developed by Iconic for the Norwegian MFA to translate legal texts from English into Norwegian. Table 23 shows the details of the 500-segment test set used for this purpose.[31]

Table 23: Size of the test set for EN→NO

***Source:* English (approximate average sentence length)**	***Target/output:* Norwegian**			
	Human reference	***Iconic***	***Google Translate***	***Microsoft Bing***
17,373 (35 tokens per segment)	15,271	15,074	14,979	14,129

31 We note that the difference in tokens between Microsoft Bing's output and the human reference is striking (14,129 vs. 15,271 tokens).

6.6.3.1 Follow-up automatic evaluation

As indicated in Table 24, Iconic's MT engine shows a very strong performance, and is the clear winner in the comparison: Google Translate and eTranslation receive similar scores, and Microsoft Bing lags far behind in fourth position.

Table 24: Results of AEMs on the 500-segment test set for EN→NO

Engine	BLEU	METEOR	TER	chrF
Iconic	48.4%	37.9	37.3%	77.0%
Google Translate	40.3%	34.1	43.7%	71.0%
Microsoft Bing	31.8%	29.6	51.2%	64.8%
eTranslation	40.8%	34.6	46.5%	70.7%

6.6.3.2 Human evaluation based on pairwise comparative ranking

The Norwegian MFA chose pairwise comparative ranking for the human evaluation, and we used the MT output from Iconic's engine and Google Translate (as the next strongest system overall, based on the AEM scores shown in Table 24) to conduct the comparison. In keeping with best practice for this type of evaluation, to avoid any bias or learning effect in the evaluators' answers due to the regular patterning in the items to be evaluated, we randomised the two outputs and presented them in a scrambled, unpredictable order to the three expert evaluators from the EA: they were shown the source/input segment in English, and next to it the outputs in Norwegian provided by the two MT systems in a random sequence. The evaluators were asked to rank the two outputs according to their preference; ties were allowed, with the options "equally good" and "equally poor". Only when the evaluation results were returned to the researchers were the outputs de-anonymised and re-attributed to their respective MT systems to analyse the results. While evaluators 1 and 2 were able to complete the evaluation of the entire sample of 500 segments, due to time constraints evaluator 3 could analyse only half of the sample, i.e., the first 250 segments, providing a total of 1,250 data points for the analysis of the pairwise comparative ranking, as shown in Table 25.

Table 25: Results of pairwise comparative ranking for the Norwegian Ministry of Foreign Affairs, EN→NO

	Evaluator 1 *(500 segments)*		**Evaluator 2** *(500 segments)*		**Evaluator 3** *(250 segments)*		**Total** **(1,250 judgements)**	
Iconic best	229	45.8%	260	52.0%	94	37.6%	583	46.6%
Google Translate best	138	27.6%	127	25.4%	68	27.2%	333	26.6%
Equally good	118	23.6%	84	16.8%	86	34.4%	288	23.0%
Equally poor	14	2.8%	29	5.8%	1	0.4%	44	3.5%
Not assigned	1	0.2%	0	0.0%	1	0.4%	2	0.1%
Total	*500*	*100%*	*500*	*100%*	*250*	*100%*	*1,250*	*99.8%*

The most frequent result of the human evaluation for all three evaluators was that Iconic's output was rated higher than Google Translate. This preference was particularly clear for evaluators 1 and 2, who preferred Iconic over Google Translate twice as many times on the complete 500-segment test set (the averaged total across the three evaluators was over 46%). Interestingly, the two outputs were deemed of equally good quality on several occasions by all three reviewers (23.6%, 16.8%, and 34.4% of the cases, respectively, i.e., 23.0% on average). By comparison, similarly poor judgements for both systems were given only infrequently by all three evaluators, for a small minority of the analysed segments. Evaluators 1 and 3 missed one evaluation item each (shown as "not assigned" in Table 25) and, due to the rounding of the percentages to one decimal place, the average combined total over the 1,250 data points was just under 100%.

6.6.4 Evaluation for the Translation Centre of Iceland's Foreign Ministry, EN→IS

Table 26 shows the statistics for the test set used to evaluate the customised MT engine in the legal domain for the Translation Centre of Iceland's Foreign Ministry.

Table 26: Details and size (in words) of the test set for EN→IS

***Source:* English (approximate average sentence length)**	***Target/output:* Icelandic**			
	Human reference	***Iconic***	***Google Translate***	***Microsoft Bing***
8,544 (17 words per segment)	7,519	7,459	7,208	7,638

6.6.4.1 Follow-up automatic evaluation

All the AEMs indicated that Iconic's engine clearly outperformed all the free online MT systems, as shown in Table 27. Google Translate and eTranslation performed similarly for the language pair EN→IS, remaining a distant second from the quality achieved by the bespoke NMT engine, while Microsoft Bing lagged some distance behind all the other systems.

Table 27: Results of AEMs on the 500-segment test set for EN→IS

	BLEU	**METEOR**	**TER**	**chrF**
Iconic	35.9%	30.5	53.4%	70.1%
Google Translate	23.6%	24.1	63.1%	58.7%
Microsoft Bing	18.4%	21.4	69.6%	52.0%
eTranslation	23.2%	24.3	69.5%	60.2%

6.6.4.2 Human evaluation based on post-editing

In line with best practices, we adopted an experimental protocol when conducting a human evaluation of time and productivity gains associated with post-editing raw MT output from Iconic's MT engine. Since two translators from the Translation Centre of Iceland's Foreign Ministry were available to act as evaluators, we swapped their roles during the evaluation, i.e., each worked in both modes of operation (MT post-editing and "standard" translation) on half of the test set, changing roles/modes of operation at the midway point, in separate sessions. We counterbalanced their roles to control for individual variables such as personal work speed, familiarity with the subject matter, inclination to adopting MT, and full post-editing (for top publishable quality) as a workflow component.

Table 28: Comparison of MT post-editing and normal translation for the Translation Centre of Iceland's Foreign Ministry, EN→IS

	Mode of operation	**No. of segments**	**Segment number**	**Overall time**
Evaluator 1	Translation	500	1–250	7h 45m
Evaluator 2	Post-editing	500	1–250	3h 28m
Evaluator 1	Post-editing	500	251–500	8h 12m
Evaluator 2	Translation	500	251–500	6h 10m

The results presented in Table 28 are inconclusive, due to very different performance on the two samples and for the two setups. Post-editing significantly helped productivity and produced clear time gains for segment numbers 1–250, but required more time for segments 251–500. We therefore note the need for closer investigation of this situation, particularly at the segment level, to check the consistency and actual gains of MT post-editing as opposed to "standard" translation, when the translators worked using the typical tools and resources available to them. Such an investigation could help to understand, for example, the impact of outliers on the overall evaluation results. This less than clear scenario is consistent with prior studies such as Plitt and Masselot (2010), Koehn and Germann (2014), and Läubli et al. (2019), which considered the variation in speed and productivity of human translators involved in post-editing tasks, and showed that drawing firm conclusions from the analysis of post-editing timing can be quite difficult. In our case, this difficulty was compounded by the fact that only two professionals participated in this part of the evaluation.

In addition, due to the software available to perform this evaluation, only the overall timing of the tasks was available (time spent on breaks, interruptions, etc., was carefully documented by the two evaluators and duly factored out from the data included in Table 28). Unfortunately, it was not possible to extract and compare in detail the time spent by the two evaluators at the segment level when translating and post-editing. The learning curve of post-editing is likely to have played a role here, as the more post-editing experience the translators acquire, the faster they tend to identify issues in the raw MT output (especially output from the specific MT engine they work with on a regular basis, once they learn its main strengths and weaknesses) and are therefore able to fix these errors more efficiently and quickly.

6.7 Discussion of the evaluation results and general observations

Overall, the evaluation of the bespoke NMT engines for the legal domain built for PRINCIPLE by Iconic showed strong results, and the human evaluations performed on the basis of the use cases of interest to the EAs generally confirmed the positive indications of the AEMs. The flexibility offered to the EAs to ensure that the evaluation protocol matched their requirements and needs was a precondition for a successful and meaningful human evaluation of the MT engines in the legal domain that validated the relevance and quality of the LRs collected by the project.

The evaluation of a range of systems for various EAs and multiple language combinations revealed that there is value in combining state-of-the-art AEMs and human approaches, and we would recommend it as a sound approach for similar efforts, including when MT engine-building is part of a larger LR collection project, as in the case of PRINCIPLE. In our case, we provided depth to the automatic part of the evaluation by beginning with the baseline evaluation, which was followed by an additional automatic evaluation of the bespoke EA NMT engines on in-domain test sets withheld from the initial training data in the legal domain.

The subsequent part of the evaluation, which also included the human component, was performed on separate 500-segment test sets, provided specifically by the EAs to match their intended use cases and reflect the types of texts for which they wished to use MT. This approach proved very effective because the MT developers (as well as the EAs themselves) could rely on quantitative and qualitative data provided by tailored human evaluation, including comments and feedback on the specific strengths and weaknesses identified in the systems by the evaluators, to gradually improve the engines and gain a realistic assessment of the advantages of introducing them into the translation workflow.

7 Conclusions

7.1 Main lessons learned

The extensive MT evaluation undertaken in PRINCIPLE that has been presented in this article underscores the importance of collecting high-quality LRs for less-supported languages, especially in the legal domain. In this article we have demonstrated that fresh data sets can be leveraged to dramatically improve the quality of NMT systems and help them to outperform general-purpose, freely available MT systems, focusing on the legal domain for four under-resourced European languages in combination with English: Croatian, Irish, Norwegian, and Icelandic.

After describing the key features of the NMT systems developed for the EAs, we gave an account of the evaluation designed and conducted for the bespoke, domain-adapted MT engines developed for them. This showed that a combination of state-of-the-art AEMs and human evaluations tailored to the specific use cases for which the MT is to be deployed can serve a twofold aim, i.e., validating the quality and effectiveness of the LRs before they are further shared with the community (e.g., to improve eTranslation and be disseminated via ELRC-SHARE for various purposes), and demonstrating the usefulness of adopting MT and post-editing in professional translation scenarios.

7.2 Importance of sharing high-quality language resources in the legal domain

The work presented in this article showcases some of the main achievements of the EU-funded PRINCIPLE project, which aimed at collecting and sharing high-quality LRs for Croatian, Icelandic, Irish, and Norwegian,

especially in the DSIs of eJustice and eProcurement, with particular focus on eJustice. We have illustrated how building and subsequently evaluating custom NMT systems represented a crucial step in the validation process of the quality and effectiveness of the LRs collected by the project. In this way, high-quality, validated LRs were made available to the wider community for a range of applications in the fields of language technologies and natural language processing, and can be used to directly improve the quality of eTranslation, for the benefit of both EU and non-EU citizens whose languages have thus far been chronically under-served.

7.3 The impact of using high-quality machine translation for legal texts

The MT engines described and evaluated in this article have been deployed and made available to the EAs in Croatia, Ireland, Norway and Iceland. In the case of the Norwegian MFA, an active interaction between the EA and Iconic led to a retraining of the engine and the subsequent provision of a new improved version. We also retrained all the baseline engines and some of the other EA engines to ensure that high-quality MT capability was achieved in the project. At the time of writing, over one million words have been translated using the bespoke MT engines trained specifically for each EA. The evaluation has shown that good-quality, in-domain training data leads to high-quality MT engines that can in turn positively impact professional translation workflows for legal texts. The success of these bespoke MT engines also validates the quality and usefulness of the digital LRs collected and shared by PRINCIPLE for Croatian, Icelandic, Irish, and Norwegian, as a valuable contribution to mitigating the adverse effects of the traditional shortage of digital data sets for these under-resourced European languages.

Acknowledgements

The PRINCIPLE project was co-financed by the European Union Connecting Europe Facility under Action 2018-EU-IA-0050 with grant agreement INEA/CEF/ICT/A2018/1761837. The authors affiliated with the ADAPT SFI Research Centre also acknowledge the financial support of Science Foundation Ireland through the SFI Research Centres Programme under Grant Agreement No. 13/RC/2106_P2. The authors are grateful to the anonymous reviewers for their insightful comments and valuable suggestions on a previous version of this paper, and to the guest editors of the special issue and the journal editors for helpful advice in the preparation of the final version of the paper. Any errors remain the sole responsibility of the authors.

References

Bago, Petra, Dunne, Jane, Gaspari, Federico, Kåsen, Andre, Kristmannsson, Gauti, McHugh, Helen, Olsen, Jon Arild, Sheridan, Dana D., Sheridan, Páraic, Tinsley, John, & Way, Andy. (2020). Progress of the PRINCIPLE project: Promoting MT for Croatian, Icelandic, Irish and Norwegian. In André Martins, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Turchi Marco, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, & Mikel L. Forcada (Eds.), *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation* (pp. 465–466). European Association for Machine Translation.

Bago, Petra, Castilho, Sheila, Dunne, Jane, Gaspari, Federico, Kåsen, Andre, Kristmannsson, Gauti, Olsen, Jon Arild, Resende, Natalia, Rúnar, Gíslason Níels, Sheridan, Dana D., Sheridan, Páraic, Tinsley, John, & Way, Andy. (2022). Achievements of the PRINCIPLE project: Promoting MT for Croatian, Icelandic, Irish and Norwegian. In Lieve Macken, Andrew Rufener, Joachim Van den Bogaert, Joke Daems, Arda Tezcan, Bram Vanroy, Margot Fonteyne, Loïc Barrault, Marta R. Costa-Jussà, Ellie Kemp, Spyridon Pilos, Christophe Declercq, Maarit Koponen, Mikel L. Forcada, Carolina Scarton, & Helena Moniz (Eds.), *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation* (pp. 349–350). European Association for Machine Translation.

Banerjee, Satanjeev, & Lavie, Alon. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, & Clare Voss (Eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization* (pp. 65–72). European Association for Machine Translation.

Bojar, Ondřej, Graham, Yvette, & Kamran, Amir. (2017). Results of the WMT17 Metrics Shared Task. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, & Julia Kreutzer (Eds.), *Proceedings of the Second Conference on Machine Translation* (pp. 489–513). European Association for Machine Translation. https://doi.org/10.18653/v1/W17-4755

Bowker, Lynne, & Ciro, Jairo Buitrago. (2019). *Machine translation and global research: Towards improved machine translation literacy in the scholarly community*. Emerald Publishing.

Callison-Burch Chris, Osborne, Miles, & Koehn, Philipp. (2006). Re-evaluating the role of BLEU in machine translation research. In Diana McCarthy, & Shuly Wintner (Eds.), *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 249–256). European Association for Machine Translation.

Castilho, Sheila, Doherty, Stephen, Gaspari, Federico, & Moorkens, Joss. (2018). Approaches to human and machine translation quality assessment. In Joss Moorkens, Sheila Castilho, Federico Gaspari, & Stephen Doherty (Eds.), *Translation quality assessment: From principles to practice* (pp. 9–38). Springer. https://doi.org/10.1007/978-3-319-91241-7_2

Chen, Boxing, Cherry, Colin, Foster, George, & Larkin, Samuel. (2017). Cost weighting for neural machine translation domain adaptation. In Thang Luong, Alexandra Birch, Graham Neubig, & Andrew Finch (Eds.), *Proceedings of the First Workshop on Neural Machine Translation* (pp. 40–46). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3205

Chu, Chenhui, Dabre, Raj, & Kurohashi, Sadao. (2017). An empirical comparison of domain adaptation methods for neural machine translation. In Regina Barzilay, & Min-Yen Kan (Eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 385–391). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2061

Chu, Chenhui, & Wang, Rui. (2018). A survey of domain adaptation for neural machine translation. In Emily M. Bender, Leon Derczynski, & Pierre Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1304–1319). Association for Computational Linguistics.

Chu, Chenhui, Dabre, Raj, & Kurohashi, Sadao. (2018). A comprehensive empirical comparison of domain adaptation methods for neural machine translation. *Journal of Information Processing*, *26*, 529–538. https://doi.org/10.2197/ipsjjip.26.529

Clark, Jonathan H., Dyer, Chris, Lavie, Alon, & Smith, Noah A. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In Dekang Lin, Yuji Matsumoto, & Rada Mihalcea (Eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 176–181). Association for Computational Linguistics.

Denkowski, Michael, & Lavie, Alon. (2012). Challenges in predicting machine translation utility for human post-editors. *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers* (article 6). Association for Machine Translation in the Americas.

Eide, Kristine, Kåsen, Andre, & Dale, Ingerid Løyning. (2022). *D1.26: Report on the Norwegian language*. European Language Equality Project.

Etchegoyhen, Thierry, Fernández Torné, Anna, Azpeitia, Andoni, Martínez García, Eva, & Matamala, Anna. (2018). Evaluating domain adaptation for machine translation across scenarios. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, & Takenobu Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 6–15). European Language Resources Association.

Gaspari, Federico, Way, Andy, Dunne, Jane, Rehm, Georg, Piperidis, Stelios, & Giagkou, Maria. (2022). *D1.1: Digital language equality (preliminary definition)*. European Language Equality Project.

Gupta, Rohit, Lambert, Patrik, Nath Patel, Raj, & Tinsley, John. (2019). Improving robustness in real-world neural machine translation engines. In Mikel L. Forcada, Andy Way, John Tinsley, Dimitar Shterionov, Celia Rico, & Federico Gaspari (Eds.), *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks* (pp. 142–148). European Association for Machine Translation.

Kahaner, Steven. (2005). Issues in legal translation. *Ccaps Translation and Localization*, 1–3.

Klubička, Filip, Kasunić, Lorena, Blazsetin, Danijel, & Bago, Petra. (2022). Challenges of building domain-specific parallel corpora from public administration documents. In Reinhard Rapp, Pierre Zweigenbaum, & Serge Sharoff (Eds.), *Proceedings of the 15th Workshop on Building and Using Comparable Corpora (BUCC 2022) within LREC 2022* (pp. 50–55). Association for Computational Linguistics.

Koehn, Philipp, & Ulrich, Germann. (2014). The impact of machine translation quality on human post-editing. In Ulrich Germann, Michael Carl, Philipp Koehn, Germán Sanchis-Trilles, Francisco Casacuberta, Robin Hill, & Sharon O'Brien (Eds.), *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation* (pp. 38–46). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-0307

Lavie, Alon, & Agarwal, Abhaya. (2007). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In Philipp Koehn, & Christof Monz (Eds.), *Proceedings of the Workshop on Statistical Machine Translation* (pp. 228–231). Association for Computational Linguistics. https://doi.org/10.3115/1626355.1626389

Läubli, Samuel, Amrhein, Chantal, Düggelin, Patrick, Gonzalez, Beatriz, Zwahlen, Alena, & Volk, Martin. (2019). Post-editing productivity with neural machine translation: An empirical assessment of speed and quality in the banking and finance domain. In Mikel Forcada, Andy Way, Barry Haddow, & Rico Sennrich (Eds.), *Proceedings of Machine Translation Summit XVII: Research Track* (pp. 267–272). Association for Computational Linguistics.

Lommel, Arle, Uszkoreit, Hans, & Burchardt, Aljoscha. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: Tecnologies de la Traducció*, *12*, 455–463. https://doi.org/10.5565/rev/tradumatica.77

Lynn, Teresa. (2022). *D1.20: Report on the Irish language*. European Language Equality Project.

O'Brien, Sharon, & Ehrensberger-Dow, Maureen. (2020). MT literacy – A cognitive view. *Translation, Cognition & Behavior*, *3*(2), 145–164. https://doi.org/10.1075/tcb.00038.obr

Papineni, Kishore, Roukos, Salim, Ward, Todd, & Zhu, Wei-Jing. (2002). BLEU: A method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, & Dekang Lin (Eds.), *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics.

Plitt, Mirko, & Masselot, François. (2010). A productivity test of statistical machine translation post-editing in a typical localisation context. *The Prague Bulletin of Mathematical Linguistics*, *93*, 7–16.

Popović, Maja. (2015). ChrF: Character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, & Pavel Pecina (Eds.), *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT–15)* (pp. 392–395). Association for Computational Linguistics.

Post, Matt. (2018). A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, & Karin Verspoor (Eds.), *Proceedings of the Third Conference on Machine Translation (WMT): Research Papers* (pp. 186–191). Association for Computational Linguistics.

Radding, Charles, & Ciaralli, Antonio. (2006). *The corpus iuris civilis in the Middle Ages: Manuscripts and transmission from the sixth century to the juristic revival*. Brill.

Rögnvaldsson, Eiríkur. (2022). *D1.19: Report on the Icelandic language*. European Language Equality Project.

Sennrich, Rico, Haddow, Barry, & Birch, Alexandra. (2016). Neural machine translation of rare words with subword units. In Katrin Erk, & Noah A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1715–1725). Association for Computational Linguistics.

Snover, Matthew, Dorr, Bonnie, Schwartz, Richard, Micciulla, Linnea, & Makhoul, John. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: "Visions for the Future of Machine Translation"* (pp. 223–231). Association for Machine Translation in the Americas.

Tadić, Marko. (2022). *D1.7: Report on the Croatian language*. European Language Equality Project.

Thompson, Brian, & Koehn, Phillip. (2019). Vecalign: Improved sentence alignment in linear time and space. In Sebastian Padó, & Ruihong Huang (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP–IJCNLP)* (pp. 1342–1348). Association for Computational Linguistics.

Varga, Dániel, Németh, László, Halácsy, Péter, Kornai, András, Trón, Viktor, & Nagy, Viktor. (2005). Parallel corpora for medium density languages. In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, & Nikolai Nikolov (Eds.), *International conference on recent advances in natural language processing (RANLP) 2005* (pp. 590–596). Association for Computational Linguistics.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, & Polosukhin, Illia. (2017). Attention is all you need. In Isabelle Guyon, Ulrike Von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, Shri N. S. Vishwanathan, & Roman Garnett (Eds.), *NeurIPS Processing: Advances in Neural Information Processing Systems 30 (NIPS 2017)* (pp. 5998–6008). Conference on Neural Information Processing Systems.

Way, Andy. (2018). Quality expectations of machine translation. In Joss Moorkens, Sheila Castilho, Federico Gaspari, & Stephen Doherty (Eds.), *Translation quality assessment: From principles to practice* (pp. 159–178). Springer. https://doi.org/10.1007/978-3-319-91241-7_8

Way, Andy, & Gaspari, Federico. (2019). PRINCIPLE: Providing resources in Irish, Norwegian, Croatian and Icelandic for the purposes of language engineering. In Mikel Forcada, Andy Way, John Tinsley, Dimitar Shterionov, Celia Rico, & Federico Gaspari (Eds.), *Proceedings of Machine Translation Summit XVII, Volume 2: Translator, Project and User Tracks* (pp. 112–113). European Association for Machine Translation.

Way, Catherine. (2016). The challenges and opportunities of legal translation and translator training in the 21st century. *International Journal of Communication*, *10*, 1009–1029.

Wolff, Leon. (2011). Legal translation. In Kirsten Malmkjær, & Kevin Windle (Eds.), *The Oxford handbook of translation studies* (pp. 228–242). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199239306.013.0017

Zhang, Mike, & Toral, Antonio. (2019). The effect of translationese in machine translation test sets. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, & Karin Verspoor (Eds.), *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)* (pp. 73–81). Association for Computational Linguistics.