An investigation of English-Irish machine translation and associated resources

Meghan Dowling B.A. (Mod.)

Supervised by Prof. Andy Way and Dr. Teresa Lynn

A thesis presented for the degree of Doctor of Philosophy

School of Computing

Dublin City University

January 5, 2022

Dedication

For Nanny and Po-Po.

Declaration

I hereby certify that this material, which I now submit for assessment on the program of study leading to the award of Ph.D. is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:

ID no: 16213542

Date: January 5, 2022

Acknowledgements

Firstly, I would like to thank my supervisors, Prof. Andy Way and Dr. Teresa Lynn, for their advice and encouragement, particularly during the final year. I thank the examiners, Prof. Mikel Forcada and Prof. Sharon O'Brien, for their helpful and insightful comments.

Thank you to my colleagues for the continued support, especially Eva Vanmassenhove, Daria Dzendzik and Abigail Walsh agus na gaeilgals ar fad - go raibh míle maith agaibh, ní bheinn in ann é seo a dhéanamh mura mbeadh bhur tacaíocht agam.

Thank you to Mam and Dad, my brothers Patrick, Seán and Joe, and my extended family and friends for always being there for a distraction, cupán tae, shoulder to cry on and the encouragement to keep going, fiú nuair a raibh mé ag snámh in aghaidh easa.

Contents

List of Figures

List of Tables

An investigation of English-Irish machine translation and associated resources

Meghan Dowling

Abstract

As an official language in both Ireland and the European Union (EU), there is a high demand for English-Irish (EN-GA) translation in public administration. The difficulty that translators currently face in meeting this demand leads to the need for reliable domain-specific user-driven EN-GA machine translation (MT). This landscape provides a timely opportunity to address some research questions surrounding MT for the EN-GA language pair.

To this end, we assess the corpora available for training data-driven MT systems, including publicly-available data, data collected through EU-supported data collection efforts and web-crawling, showing that though Irish is a low-resource language it is possible to increase the corpora available through concerted data collection efforts. We investigate how increased corpora affect domain-specific (public administration) statistical MT (SMT) and neural MT (NMT) systems using automatic metrics. The effect that different SMT and NMT parameters have on these automatic values is also explored, using sentence-level metrics to identify specific areas where output differs greatly between MT systems and providing a linguistic analysis of each.

With EN-GA SMT and NMT automatic evaluation scores showing inconclusive results, we investigate the usefulness of EN-GA hybrid MT through the use of monolingual data as a source of artificial data creation via backtranslation. We evaluate these results using automatic metrics and linguistic analysis. Although results indicate that the addition of artificial data did not have a positive impact on EN-GA MT, repeated experiments involving Scottish Gaelic show that the method holds promise, given suitable conditions.

Finally, given that the intended use-case of EN-GA MT is in the workflow of a professional translator, we conduct an in-depth human evaluation study for EN-GA SMT and NMT, providing a human-derived assessment of EN-GA MT quality and comparison of EN-GA SMT and NMT. We include a survey of translator opinions and recommendations surrounding EN-GA SMT and NMT as well as an analysis of data gathered through the post-editing of MT output. We compare these results to those generated automatically and provide recommendations for future work on EN-GA MT, in particular with regards to its use in a professional translation workflow within public administration.

Chapter 1

Introduction

The Irish language is the national language of Ireland. It is the first official language of Ireland and an official language of the European Union (EU). It is also considered a minority language, sharing official status in Ireland with English, the majority language.

Although the majority of Irish speakers are fluent in English (Higgins and Ní Uigín, 2017),[1] Irish speakers have the right to access public information and services through their native language. This right is made concrete via legislation, both at a national and European level. Within the Republic of Ireland, the Official Languages Act (2003) requires official public information and services to be available in both Irish and English.[2]

However, despite the legal obligation for the production of Irish language content nationally, in practice the Official Languages Act (2003) applies only to certain, and not all, public content:

> *"The direct provisions of Section 10 of the Act deal with the duties of public bodies to publish certain documents simultaneously in both official languages, for example documents containing public policy proposals, annual reports, financial statements and specific strategy statements."* An

[1] Although to date, no studies have provided the number of Irish speakers who do not speak Irish, Riagáin (2017) estimates that no more than 5% of the population in the Republic 'use Irish as their first or main language.'

[2] `http://www.irishstatutebook.ie/eli/2003/act/32/enacted/en/html`

> Coimisinéir Teanga (2003).

Within the EU, there became an obligation to provide Irish language content in 2005 when Irish became an official EU language. However, the obligation to produce Irish language text relates only to a limited amount of content, due to the derogation that is currently placed on the production of Irish language content (Publications Office of the European Union, 2011).

> "*Irish has been an EU language since 2005 with a limited status, meaning that only a small share of documents were translated into Irish*" (The European Commission, 2020).

Although only a portion of all national and EU public documents are being made available in Irish, at present the demand for bilingual content exceeds the productivity capabilities of translation services in Irish government departments and the EU and, despite concerted recruitment efforts, cannot be fully met by human translators alone (Lynn et al., 2019).

> *"EPSO [European Personnel Selection Office] competitions for GA translators and linguistic assistants were launched in 2016. The translator competition attracted 210 applicants and yielded 10 successful candidates against a target of 62. Some 8 of 10 were already employed by the institutions on temporary contracts, so the net gain in capacity was 2."* The European Commission (2019).

The discrepancy between demand for GA content and available professional GA translators will become more severe once the derogation granted to the Irish language runs out at the end of 2021. Accordingly, we contend that this increasing imbalance between supply and demand necessitates a technology-orientated solution, which we explore in this thesis.

One such technology-oriented solution is the introduction of machine translation (MT) as a productivity tool for professional translators. MT is the attempted translation of natural languages via computational means. The MT systems described in this work are trained using corpus-based approaches, i.e. bilingual text (also known as parallel corpora), and sometimes monolingual text, are used to train MT systems (see Chapter 3 for a detailed explanation of the most common MT paradigms).

> *"... previously unseen texts are automatically translated using information gleaned from examples of past translations produced by humans"*
> Hearne and Way (2011, p. 1).

MT is well-established in the workflow of a professional translator, whereby translators may have the option to post-edit machine-translated text rather than translating from scratch (Moorkens et al., 2018). However, the limited amount of Irish translations being produced has led to Irish being classified as a less-resourced language in terms of MT (Judge et al., 2012). As a less-resourced and minority language, the Irish language has not enjoyed the benefits of technological advancements in the field of MT to the same extent that well-resourced languages (such as English) have.

The aim of this work is to improve English–Irish (EN–GA) MT so that it may be used as a practical aid in the production of bilingual text at a national and European level. If done properly, EN–GA MT will be invaluable in meeting the language rights needs of Irish speakers.

We believe that a 'virtuous cycle' exists in the relationship between MT and translation production/data gathering (see Figure 1.1). The better the MT system is at producing candidate translations that are useful for professional translators, the more productive these translators can be. Greater productivity in turn leads to lower costs and more bilingual output. This bilingual output can then be used to improve the MT systems, which will lead to better productivity, and so on, continuing the virtuous cycle.

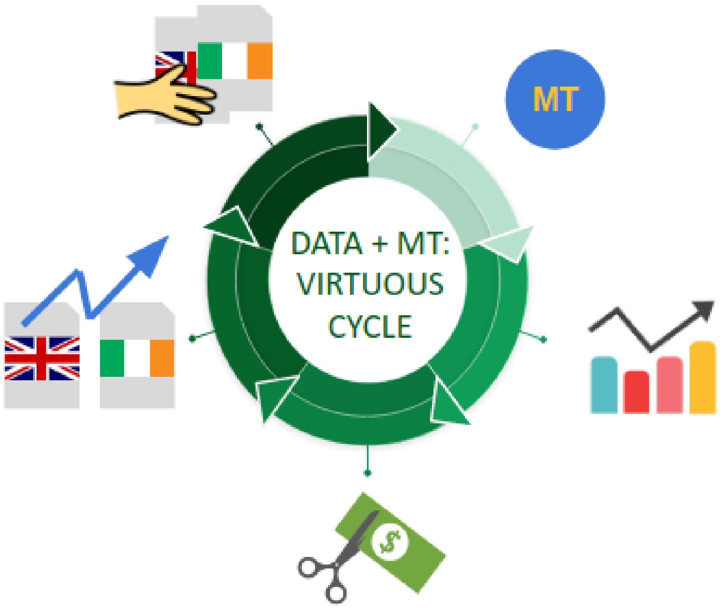


Figure 1.1: Virtuous cycle of MT and data collection

We believe that language technology resources are vital for the preservation and growth of every language and that it is necessary to develop methods of creating MT systems for languages without an extensive amount of language data available.

Our research focuses on examining a number of possible ways to improve EN–GA MT. These approaches include gathering and curating suitable datasets, experimenting with MT infrastructures and also testing the use of various datasets.

1.1 Overview of English–Irish machine translation

Although EN–GA MT has been a neglected area of MT research, there are a number of examples of EN–GA MT systems. In this section, we list some other known EN–GA MT systems and their uses.

1.1.1 The Tapadóir project

While EN–GA MT is rarely used for comprehension purposes, the primary focus of the application of Irish MT is within the context of a professional translation workflow (involving post-editing by human translators). Until recently, however, MT was not used as a translation tool at an official national level (Lynn et al., 2019).

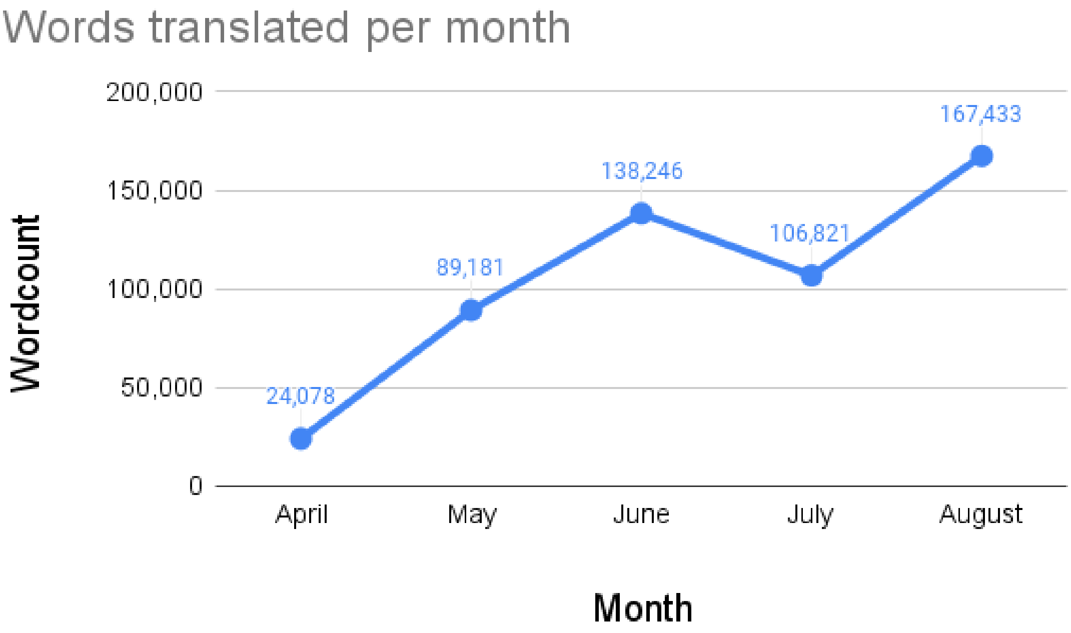


Figure 1.2: Graph showing the number of words the MT system was used to translate during the first 5 months of the Tapadóir project.

The Tapadóir project was a pilot study which aimed to investigate the addition of MT in the translator workflow of an official Irish government department.[3] This study explored whether one method commonly used to improve the quality of low-resource MT is to focus on tailoring a system to a particular domain. The Tapadóir project (Dowling et al., 2015) has shown that the translation of EN–GA documents could be aided by a domain-tailored statistical machine translation (SMT) system. This exploratory study into introducing MT into the workflow of English–Irish translators within an Irish government department has shown that the development of EN–GA MT is possible, yet contains scope for improvement. Figure 1.2 shows a graph of the uptake on the Tapadóir MT plug-in in the DCHG during the first 5 months.[4] This was the first time that MT was used in an official in-house translation

[3] The Department of Culture, Heritage and the Gaeltacht (DCHG). DCHG is the Irish government department responsible for ensuring that the Irish-language needs of the Irish public are being met by the government: `https://www.chg.gov.ie/`

[4] Note that the decrease in numbers in July can be attributed to the Irish government taking a scheduled annual break.

capacity within an Irish government department with translators given the choice to use MT, a translation memory (TM) fuzzy match, or translate from scratch. Uptake in use of the MT output was gradual, increasing as the translators became more used to post-editing. In this thesis, we build on the resources and insights gained during the Tapadóir project, maintaining a working relationship with DCHG.

1.1.2 eTranslation

At a European level, eTranslation[5] is a multilingual MT system built for translating between official EU languages. It is used in the workflow of professional translators within the European Commission, and can also be used by those working in an official public capacity.

> *"The eTranslation service provides the ability to translate formatted documents and plain text between any pair of EU official languages, as well as Icelandic, Norwegian (Bokmål) and Russian (EN<->RU), while preserving to the greatest extent possible the structure and format of those documents. CEF eTranslation builds on the European Commission's earlier machine translation service, MT@EC, which was developed by the Directorate-General for Translation under the Interoperability Solutions for European Public Administrations (ISA) programme."* Connecting Europe Facility (2020)

1.1.3 Publicly-available online interfaces

MT tools accessed via online interfaces are not always viable solutions for public bodies due to privacy concerns,[6] and as such are most commonly used by the general public to provide a quick rough translation, e.g. on social media (Lohar, 2020). Although EN–GA MT is not usually used for gisting purposes, it is a possible use-

[5] `https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/eTranslation`

[6] Not all information translated by public bodies is intended for public consumption, e.g. internal reports, etc.

case for the opposite language direction (GA→EN MT). There are a number of online MT interfaces available for this purpose:

- Microsoft translator [7]
- Google translate [8]
- Iris (no longer live) (Arcan et al., 2016)

1.2 Features of the Irish language that can pose a challenge for machine translation

In addition to being a low-resource language, there are some linguistic features of the Irish language that pose a challenge for MT. In this section we outline some of the linguistic differences between English and Irish that may present a challenge for English–Irish MT.

1.2.1 Irish language features partly addressed in this thesis

One feature of Irish that can have an effect on EN–GA MT quality is its inflected nature. Irish words can inflect for number, tense, person, case, mood and gender. Some ways that Irish words inflect include lenition (the infixing of a 'h' after the first consonant, or the prefix 't' added to the beginning of a word, e.g. 'an tsráid'), eclipsis (a type of initial mutation where a letter is added to the beginning of the word) and slenderisation (changing a 'broad' vowel ('a', 'o' or 'u') to a 'slender' vowel ('i' or 'e') or a).[9] A typical example of noun inflection can be seen in Example 1.1, using the feminine noun 'beach', meaning bee.[10] This can lead to data sparsity wherein inflected words are seen infrequently in the training data and the incorrect

[7] `https://www.microsoft.com/en-us/translator/blog/2020/01/23/dia-daoibh-ta-gaeilge-againn/`

[8] https://translate.google.com/

[9] Slenderisation can also affect consonants, for example when suffixes are changed from broad to slender, e.g. '*taoiseach*' to '*taoisigh*' and can add vowels to words, e.g. in the case of '*beiche*', seen in Example 1.1

[10] For clarity, the inflection markers (letters) in each example are displayed in bold

inflection is often produced in the MT output. Inflection (b**h**each), eclipsis (**m**beach) and slenderisation (be**i**ch**e**) can all be seen in this example. In this example, 'bee' or 'bees' in English could be translated in 5 different ways in Irish, depending on the context. This is called 'one-to-many translation' where the MT system is expected to learn many possible translations for one input. This is another example of data sparsity in EN–GA MT, where some but not all of the Irish forms may be present in the training data.

beach	bee/a bee
an b**h**each	the bee
beach**a**	bees
dath na be**i**ch**e**	the colour of the bee
dath na **m**beach	the colour of the bees

Table 1.1: The Irish noun '*beach*' inflecting in different contexts.

Inflection can also be found in Irish verbs. The example in Table 1.2 shows the regular verb *ceannaigh*, 'to buy', inflecting for person and tense (in this case, the conditional mood). In this example, 'would buy' in English could be translated in 5 different ways in Irish, depending on the context. This is called a 'one-to-many translation' where the MT system is expected to learn many possible translations for one input. This is another example of the aspect of data sparsity in EN–GA MT, where some but not all of the Irish forms may be present in the training data. One possible way to minimise the effects of cases like this would be to use byte pair encoding (BPE; see Sennrich et al. (2016a) and a further discussion in Chapter 3).

Cheann**óinn**	I would buy
Cheann**ófá**	You would buy
Cheann**ódh** sé/sí	He/she would buy
Cheann**óimis**	We would buy
Cheann**ódh** sibh	You (plural) would buy
Cheann**óidís**	They would buy

Table 1.2: The Irish verb '*ceannaigh*' inflecting for person in the conditional mood.

Another challenge for building EN–GA MT systems is the divergent word order between English and Irish. Irish follows a verb–subject–object (VSO) sentence structure, differing from the subject-verb-object (SVO) structure of English, as illustrated in Figure 1.3.

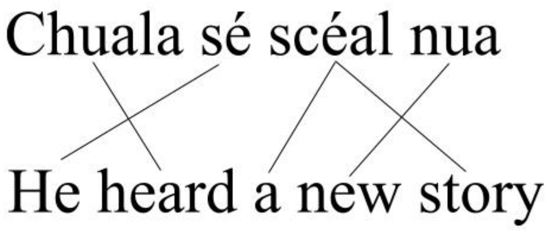


Figure 1.3: An example sentence highlighting the divergent word order between Irish and English

While a short sentence such as that in Figure 1.3 may not pose a problem for MT systems, this difference in sentence structure can negatively impact MT quality when translating longer or more complicated sentences (Koehn and Knowles, 2017). The Example in Table 1.3 shows a sentence similar to that in Table 1.3, but slightly longer. Already it can be seen that the position of the verb *chuala*, 'heard' is further away from the original position in the sentence, and therefore could pose an issue when translating.

GA sentence:	**Chuala**	an	fear	leis	an	bhféasóg	scéal	nua.
EN gloss:	**heard**	the	man	with	the	beard	story	new.
EN sentence:	'*The man with the beard **heard** a new story.*'							

Table 1.3: An Irish (GA) sentence with English (EN) gloss and corresponding sentence.

1.2.2 Other divergent features of English and Irish that may pose a challenge to English–Irish machine translation

No direct translation for 'yes' or 'no' in Irish In Irish, there is no single standalone word that directly translates to 'yes' or 'no' regardless of context. Rather, the main verb in the question is repeated in the positive for yes, and the negative for no. This is illustrated in the example in Table 1.4, where the questioning verb

Ar **cheannaigh** tú mála?	Did you buy a bag?
Cheannaigh	Yes (lit. 'bought')
Níor **cheannaigh**	No (lit. 'didn't buy')

Table 1.4: An example illustrating how a question can be answered in the positive and the negative in Irish

'*cheannaigh*' is given in the positive as a translation of 'yes' ('cheannaigh') and in the negative as a translation of 'no' ('níor cheannaigh').

This challenge is exacerbated when one takes into consideration irregular verbs, where the positive and negative terms may not resemble each other (see Table 1.5).

An **bhfaca** tú mála?	Did you see a bag?
Chonaic	Yes (lit. 'bought')
Ní **fhaca**	No (lit. 'didn't buy')

Table 1.5: An example illustrating how a question can be answered in the positive and the negative in Irish

This can be expected to pose a problem for both SMT and NMT, particularly for phrase-based SMT, where phrase tables store candidate translations of phrases. Furthermore, as many MT systems translate one line of input at a time, a single 'yes' could be extremely difficult to translate without the context of the previous sentence(s). To this end, document-level MT could be a possible avenue of research, where the MT system is trained to translate entire documents, thus maintaining the context from the rest of the document (e.g. Gong et al. (2011) and Miculicich et al. (2018)).

Different words for counting people in Irish In Irish, there are different counting words depending on whether the noun in question is human or not. In Table 1.6, we see that the number in English (e.g. 'two' is translated in differing ways if the following noun is human (e.g. musicians or cows)). This can contribute to data scarcity in MT, when not all forms are present in the training data.

Other differing characteristics

- No indefinite article in Irish (e.g. '*fear*', a man, 'beach', a bee)

- Divergent order of adjective and noun (e.g. '*fear maith*', (a) good man)

Duine amháin	One person
Carr amháin	One carr
Beirt fhear	Two men
Dhá bhord	Two tables
Triúr	Three people
Trí cathaoir	Three chairs
Ceathrar ceoltóir	Four musicians
Ceithre bhó	Four cows
Cúigear	Five people
Cúig phunt	Five euros
Seisear altra	Six nurses
Sé mhí	Six months
Seachtar	Seven people
Seacht gcapall	Seven horses
Ochtar múinteoir	Eight teachers
Ocht n-asal	Eight donkeys
Naonúr	Nine people
Naoi gcat	Nine cats
Deichniúr páiste	Ten children
Deich bpeann	Ten pens

Table 1.6: Counting people versus counting objects from one to ten in Irish

We note that this is not an exhaustive list of the complex Irish features that can have an impact on MT. For more information please see Dryer and Haspelmath (2013) and Christian Brothers (1962).

1.3 Research Questions

Our work focuses around two central research questions.

Research Question 1

As Irish is a less-resourced language, large datasets are not readily available for building high-quality MT systems. Therefore, our first research question is:

RQ1: What are the existing corpora available for use in EN–GA MT? Given related projects in this area, how can we increase the size of this corpus?

This question is discussed extensively in Chapter 2.

Research Question 2

We believe that in order for MT systems to be useful for helping meet the rights and needs of Irish language speakers, they should be robust, user-orientated and fit-for-purpose. As a less-resourced language, creative methods of building MT systems should be explored, with evaluation based on both automatic and human methods.[11] Therefore, our second research question, broken down into 3 sub-questions, is: **RQ2 (a):** How can improvements in the quality of EN–GA statistical machine translation (SMT) and neural machine translation (NMT) be captured in terms of automatic evaluation metrics?

RQ2 (b): How do these SMT and NMT systems compare; do the different systems produce the same types of errors? If so, would a hybrid SMT-NMT system outperform both baselines? Would this be confirmed in experiments on a very closely-related language pair?

RQ2 (c): How can improvements in the quality of EN–GA statistical machine translation (SMT) and neural machine translation (NMT) be captured in terms of human evaluation? To what extent do the findings from the human evaluation corroborate the findings from the automatic evaluations?

These questions are explored in Chapters 3, 4 and 5.

1.4 Roadmap

This thesis is laid out as follows:

Chapter 2 takes steps towards addressing RQ1 by providing a study of publicly-available data for use in EN–GA MT, and describes how we undertook the task of increasing the size of this dataset through data-collection efforts.

Chapter 3 explores the differences between SMT and NMT, in terms of automatic evaluation and a limited linguistic analysis. We also report on the changes in automatic evaluation scores as data is added and parameters are tailored. We

[11] See Chapter 3 for a discussion about what ‘quality’ refers to in the field of MT.

show that while an out-of-the-box NMT system scores poorer than a tailored SMT system, the addition of more data and experimentation of NMT parameters provides an NMT system that outscores a tailored SMT system trained on the same data.

RQ2(b) is explored in Chapter 4. We experiment with the use of backtranslation as a method for combining SMT and NMT systems via the creation of artificial parallel data. We show that, although automatic scores indicate that backtranslation was not beneficial with the datasets and MT systems used, similar experiments involving Scottish Gaelic MT show that this method still has huge potential.

Chapter 5 explores the differences in SMT and NMT output in terms of human evaluation. This study, the first of its kind for the EN–GA pair, employs professional translators to post-edit SMT and NMT output and provide recommendations. Results, both qualitative and quantitative, indicate NMT to be the preferred paradigm of the translators involved.

Finally, Chapter 6 provides our conclusions, as well as potential avenues for future work.

1.5 Publications

This thesis is based in part on a number of peer-reviewed publications, published in workshops, conferences and journals. In these chapters, these papers were expanded upon via the addition of more experiments and a more in-depth discussion of results. The publications which form the basis of each chapter are discussed below.

1.5.1 Relation to Published work

Chapters 2 and 3

Earlier versions of the SMT/NMT experiments described in Chapter 3, as well as data collection efforts described in Chapter 2, were published in paper format and presented at the the Language & Technology Conference (LTC) workshop on Language Technologies in support of Less-Resourced Languages (LRL) 2016, the

JEP-TALN Celtic Language Technology Workshop (CLTW), the European Association for Machine Translation (EAMT) workshop on Social Media and User Generated Content Machine Translation, and the Association for Machine Translation in the Americas (AMTA) workshop Technologies for MT of Low Resource Languages (Dowling et al., 2015; Dowling et al., 2016; Dowling et al., 2018).

- Dowling, Meghan, Lauren Cassidy, Eimear Maguire, Teresa Lynn, Ankit Srivastava, and John Judge (2015). “Tapadóir: Developing a Statistical Machine Translation Engine and Associated Resources for Irish”. In: *The 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the The Fourth LRL Workshop: "Language Technologies in support of Less-Resourced Languages"*. Poznan, Poland, pp. 314–318.

- Dowling, Meghan, Teresa Lynn, Yvette Graham, and John Judge (2016). “English to Irish Machine Translation with Automatic Post-Editing”. In: *Journées d’Études sur la Parole Traitement Automatique des Langues Naturelles Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (JEP-TALN-RECITAL). The 2nd Celtic Language Technology Workshop*. Paris, France, pp. 42–54.

- Dowling, Meghan, Teresa Lynn, and Andy Way (2017). “A crowd-sourcing approach for translations of minority language user-generated content (UGC)”. in: *The 20th Annual Conference of the European Association for Machine Translation. First workshop on Social Media and User Generated Content Machine Translation*, pp. 1–12.

- Dowling, Meghan, Teresa Lynn, Alberto Poncelas, and Andy Way (2018). “SMT versus NMT: Preliminary comparisons for Irish”. In: *Technologies for MT of Low Resource Languages*. Boston, USA, pp. 12–20.

Chapter 4 Work which partly forms the basis of Chapter 4 was published in TEANGA, the Journal of the Irish Association for Applied Linguistics and was presented and published at the Machine Translation summit workshop CLTW (Dowling et al., 2019a; Dowling et al., 2019b).

- Dowling, Meghan, Teresa Lynn, and Andy Way (2019a). "Investigating back-translation for the improvement of English-Irish machine translation". In: *TEANGA, the Journal of the Irish Association for Applied Linguistics* 26, pp. 1–25.
- Dowling, Meghan, Teresa Lynn, and Andy Way (2019b). "Leveraging back-translation to improve machine translation for Gaelic languages". In: *Machine Translation Summit XVII.* vol. 604: *3rd Celtic Language Technology workshop.* Dublin, Ireland, pp. 58–62.

Chapter 5 The human evaluation study which Chapter 5 is based on was published at EAMT 2020, and will be presented on-line in November 2020 (Dowling et al., 2020).

- Dowling, Meghan, Sheila Castilho, Joss Moorkens, Teresa Lynn, and Andy Way (2020). "A human evaluation of English-Irish statistical and neural machine translation". In: *Proceedings of the 20th Annual Conference of the European Association for Machine Translation.* On-line (Lisbon), pp. 431–440.

1.5.2 Additional Publications

Other publications (Graham et al., 2016; Poncelas et al., 2018; Raghallaigh et al., 2019) that were co-authored during this PhD but are not directly related to the work conducted in this thesis are listed below:

- Graham, Yvette, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi (2016). "Is all that glitters in machine translation quality estimation really gold?" In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 3124–3134.

- Raghallaigh, Brian Ó, Kevin Scannell, and Meghan Dowling (2019). “Improving full-text search results on duchas.ie using language technology”. In: *Machine Translation Summit XVII. Proceedings of the Celtic Language Technology Workshop*. Dublin, Ireland, pp. 63–69. (Awarded best paper)

- Poncelas, Alberto, Kepa Sarasola, Meghan Dowling, Andy Way, Gorka Labaka, and Iñaki Alegria (2019). “Adapting NMT to caption translation in Wikimedia Commons for low-resource languages”. In: *Procesamiento del Lenguaje Natural* 63, pp. 33–40.

Chapter 2

Data collection

Within the setting of MT development, the term ‘data’ usually refers to parallel (bilingual) and monolingual (target language only) text which can be used in the training of MT systems.Data collection is an integral part of the development of corpus-based MT systems, especially for low-resource language pairs.

The two most dominant MT paradigms today, statistical machine translation (SMT) and neural machine translation (NMT), require a large amount of bilingual text for training systems. This is especially true for NMT, which requires more parallel data than SMT to train a model with good performance (see Figure 2.1 for an illustration of data size versus automatic evaluation scores for SMT and NMT[1]). Subsequent related work shows that an optimized NMT system can out-score a phrase-based SMT system even with a lower amount of corpora (Sennrich and Zhang, 2019). With a low-resource language such as Irish, this data is not readily available in large amounts, as it is with well-resourced majority languages such as English or German. In this chapter we outline the resources that are currently available for the EN-GA pair, describe our efforts in collecting this data as well as the ongoing data collection efforts that will continue to improve EN-GA MT resources. We discuss methods for data collection, present a new collection of language resources for use in Irish MT, and give recommendations for collecting data in a low-resource scenario.

[1] Quality for NMT starts much lower, outperforms SMT at about 15 million words, and even beats a SMT system with a big 2 billion word in-domain language model under high-resource conditions. Source: Koehn and Knowles (2017).

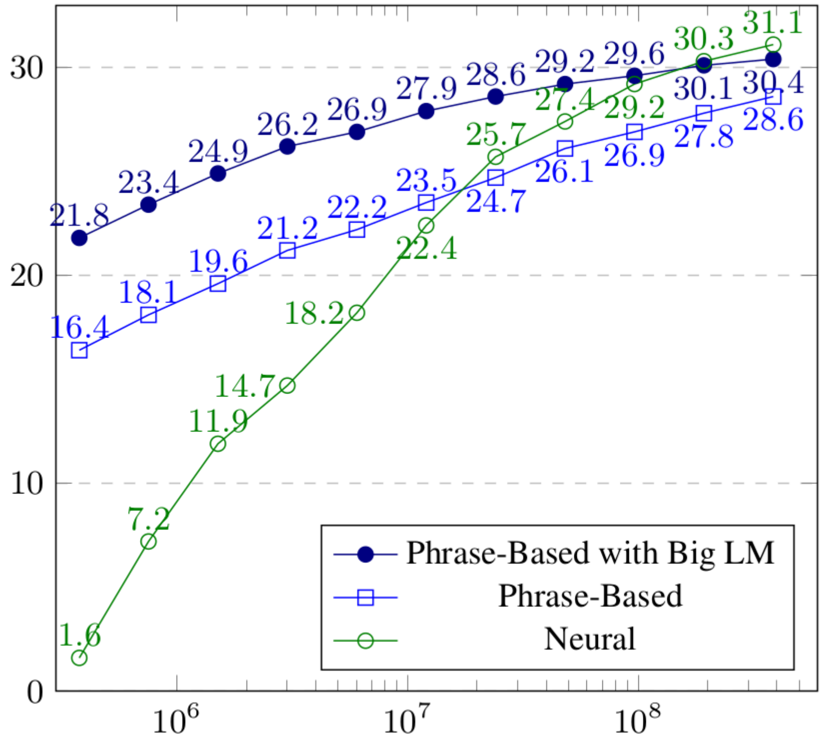


Figure 2.1: BLEU scores for English-Spanish systems trained on 0.4 million to 385.7 million words of parallel data.

2.1 Introduction

Data sparsity affects low-resource languages, and is a two-fold problem for EN-GA MT. The first type of data sparsity that affects EN-GA MT relates to the fact that Irish is more heavily inflected than English, i.e. for one Irish lemma there could be many surface forms. This can lead to a 'one-to-many' translation situation, whereby a single word in English could have many different Irish translations, depending on the context (see Table 1.1 for an example). This makes it more difficult for MT systems to 'learn' the correct translations, furthering the need for more EN-GA data (see Section 1 for a more detailed explanation). The other aspect of data sparsity pertains to the availability of language data. If the amount of data available to train the translation models is scarce, then it is much more difficult to achieve high-quality translations. This is sometimes referred to as 'data scarcity', and can be a huge hurdle for MT development involving low-resource languages. One method for dealing with data scarcity is making concerted efforts with respect to data collection.

As an official language in Ireland and the EU, one might expect a large amount of bilingual data to be available. Although there is a requirement for official public content to be made available in Irish in the Republic of Ireland,[2] in practice this requirement does not apply to all public bodies and departments, leading to a lack of publicly available information in Irish (see Chapter 1 for more information).

Compounding the problem of a lack of publicly available data, is the status quo of language data management on a national level (Lynn et al., 2019). In general, there is no culture of appointing language data officers or of having a well-developed data management system within the public service. It is common practice to outsource translations without requesting the resulting TMX files (Lynn et al., 2019). This makes it difficult to gain access to public data, especially public data that is in a suitable format. Where data is available, in general computer-assisted translation tools are not used, and it is rare for public data to be available in translation memory (TM) format.

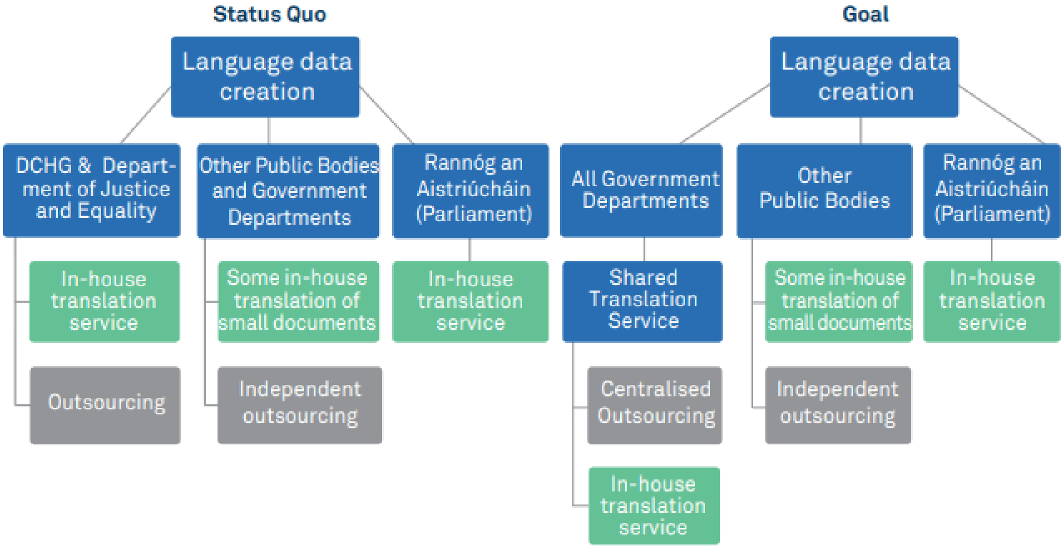

Status Quo
Language data creation
DCHG & Department of Justice and Equality
Other Public Bodies and Government Departments
Rannóg an Aistriúcháin (Parliament)
In-house translation service
Outsourcing
Some in-house translation of small documents
Independent outsourcing
In-house translation service
Goal
Language data creation
All Government Departments
Other Public Bodies
Rannóg an Aistriúcháin (Parliament)
Shared Translation Service
Centralised Outsourcing
In-house translation service
Some in-house translation of small documents
Independent outsourcing
In-house translation service

Figure 2.2: Diagram illustrating the current translation landscape in the public sector of Ireland (status quo) and a proposed workflow (goal).

"..only two of the 17 Government Departments (Department of Culture, Heritage and the Gaeltacht (DCHG) and the Department of Justice

[2] The Official Languages Act (2003) requires official public information and services to be available in both Irish and English: `http://www.irishstatutebook.ie/eli/2003/act/32/enacted/en/html`

> *and Equality) have in-house staff translators (who use computer-assisted translation (CAT) tools).. It is not common practice for public bodies or government departments to request the return of TMX files from an LSP (which is a by-product of a translation procurement)."* Lynn et al. (2019), see Figure 2.2.[3]

Figure 2.2 illustrates the current translation landscape in the public sector in the Republic of Ireland (represented under the heading 'Status Quo'), compared to the proposed improved landscape (represented under the heading 'Goal'), as reported by Lynn et al. (2019). Rather than the DCHG having its own in-house translation, and other government departments performing limited in-house translation, current plans propose that there be a shared translation service available for all government departments. In this way, departments could share translation and post-editing tools such as terminologies, translation memories and MT systems.

Within a European setting, the availability of bilingual EN-GA text is also limited. There is a derogation currently in place with respect to the production of Irish language content within the EU, meaning that relatively little Irish-language content is produced in comparison to other official EU languages.

> *"The institutions of the European Union shall not be bound by the obligation to draft all acts in Irish and to publish them in that language in the Official Journal of the European Union... but [the derogation] is to be gradually reduced in scope and eventually brought to an end by 31 December 2021."* Publications Office of the European Union (2011).

This derogation is being gradually lifted and is due to be fully lifted at the end of 2021, by which stage there will be a significant increase in the number of translators needed, compared to current requirements (see Chapter 1 for more information). While efforts are underway to increase the number of translators available (e.g. a new

[3] Source: ELRC WHITE PAPER: Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe, Lynn et al. (2019).

masters of arts (MA) in translation),[4] meeting the quota necessary to produce this volume of translations will still prove to be challenging without adequate technology. The NMT-based EN-GA eTranslation system in use by the Directorate General for Translation (DGT) is still in its early days of uptake, partly due to the lack of data available for training (CEF Digital, 2020). Unable to simply download official national or EU translation data in a suitable format for training EN–GA MT systems, other methods must be employed in order to develop the size of GA datasets.

2.2 Background and Related work

The gathering of language data for use in speech and language technologies research is vital, to the extent that there is a major conference dedicated to it: Language Resources and Evaluation Conference (LREC). In one example of a paper published in the LREC proceedings, Petukhova et al. (2012) describe the data collection and parallel corpus compilation for the creation of MT systems for use in subtitling for 14 language pairs. They cite incorrect formatting and varying file types among the common hurdles when curating their parallel corpora.

Other LREC papers discuss gathering data for which the language pairs are not low-resource, but the specific domain is. For example, Mendels et al. (2018) describe their approach to collecting code-switched English-Spanish (EN-ES) user-generated content. Although code-switching is not a topic which is dealt with in this thesis and the EN-ES pair is not usually considered low-resource, there are far fewer resources for code-switched language pairs. In this way, the use case is similar to that of a low-resource language. The authors used language identifiers to collect code-switched content from Twitter, a social media platform which can be web-crawled. Furthermore, code-switching is a phenomenon which has been shown to be present between English and Irish (Lynn and Scannell, 2019). While not addressed

[4] https://www.nuigalway.ie/courses/taught-postgraduate-courses/translation-studies.html

in this thesis, translation of code-switched EN-GA content could be an avenue for future research.

Guzmán et al. (2019) describe harnessing Wikipedia in building datasets for a low-resource language pair (Nepalese–English). They state that MT systems, particularly NMT systems, underperform when translating low-resource language pairs and discuss the importance of widespread data collection for low-resource languages.

> *"... in face of the scarcity of clean parallel data, MT systems should be able to use any source of data available, namely monolingual resources, noisy comparable data, as well as parallel data in related languages." Guzmán et al. (2019, p. 6098)*

In a similar use-case (English plus low-resource language), ShweSin et al. (2018) describe how they created a large-scale parallel corpus of Myanmar-English for use in NMT. Their dataset consists of crawled data, data available to the public, and data made available to them via NLP projects such as the Asian Language Treebank.[5]

There are a number of methods of data collection relating to MT. Crowd-sourcing – eliciting paid or un-paid participants to complete tasks such as manual translation or quality estimation – is one such method (e.g. Ambati and Vogel (2010), Chen and Dolan (2011), Aranberri et al. (2017), and Graham et al. (2017)).

> *"Crowdsourcing can allow inexpensive and rapid data collection for various NLP tasks."* Chen and Dolan (2011, p. 190–200).

Web crawling is also a common method of the collection of data for building MT systems (e.g. Rubino et al. (2015), Toral et al. (2017), Esplà-Gomis et al. (2019), and Wenzek et al. (2020)). Web crawling involves the scraping of monolingual or bilingual text from websites known to contain data in the desired language(s).

> *"The vast amount of texts publicly available in many languages has lead to a view of the web as a huge corpus that can be... exploited in applied research fields like MT."* Toral et al. (2017, p. 1019–1051).

[5] `https://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/`

These methods can also be combined, as in Toral et al. (2017), where crowd-sourcing and web crawling are combined to build parallel corpora in an efficient way. Whichever method is used to increase MT corpora, the first step should always be to evaluate the existing landscape of publicly available data.

Extensive data collection is not a novel concept for researchers of less-resourced languages. Prys and Jones (2018) provide a case study on collecting Welsh language resources for use in speech technology. Welsh belongs to the same language family as Irish and shares traits in common with Irish, both in linguistic and social terms. The case study reported how crowd-sourcing, combined with involvement in organised data gathering projects such as Mozilla's CommonVoice[6] project was used to supplement existing data from Welsh Wicipedia.[7] They also provide recommendations for researchers of low-resource or minority languages:

> *"Combining use of local, language-specific knowledge and resources with global tools and initiatives can provide the help needed to level the playing field for digitally excluded language communities, and such combinations are to be welcomed."*

Scottish Gaelic, another Celtic language, has a project dedicated to the building of language data: The Digital Archive of Scottish Gaelic/Dachaigh airson Stòras na Gàidhlig (DASG)(Ó Maolalaigh, 2016).[8] DASG aims to provide a digital corpus of monolingual data (Corpus na Gàidhlig) and a Scottish Gaelic-English dictionary (Faclair na Gàidhlig), using a team of linguists, developers and corpus experts.

While the premise of 'the more data the better' can be argued for both SMT and NMT development, we look to prominent research on low-resource MT for indications of how much data we should aim to collect. In Koehn and Knowles (2017), NMT scores only surpass those of SMT when over 100 million words of parallel data is used in training data. Sennrich and Zhang (2019), however, show

[6] More details of this project can be found here: `https://voice.mozilla.org/`

[7] `https://cy.wikipedia.org/`

[8] `https://dasg.ac.uk/`

that NMT can outscore SMT with a much smaller parallel corpus – using a Korean-English corpus of 2.3 million words as a case study. Semi-supervised (e.g. Kocmi and Bojar (2018)) and unsupervised NMT (e.g. Artetxe et al. (2017)) methods have also been reported as achieving competitive scores, though they still operate under the assumption that some other data requirements are met, e.g. the availability of monolingual data or parallel data from similar language pairs. We summarise the data used in these experiments in Table 2.1.[9]

Authors	**Language pair**	**Type**	**No. of words**
Koehn and Knowles (2017)	Spanish–English	Parallel	100 million
Sennrich and Zhang (2019)	Korean–English	Parallel	2.3 million
Kocmi and Bojar (2018)	Finnish–English	Parallel (related lang.)	44 million
Artetxe et al. (2017)	German–English	Monolingual (German)	0.9 billion words

Table 2.1: Corpora of prominent MT papers on low-resource settings with the language (lang.) pair, type of data and number (no.) of words.

Though none of the experiments collated in Table 2.1 use English-Irish as a language pair, if we use the findings as a rough guideline it would indicate that we should aim to collect between 2.3 million and 100 million words of parallel data. Therefore we propose that collecting 2.3 million words of parallel data should be our primary target, with 100 million words as the ultimate goal. We posit that other types of data such as monolingual data, though perhaps not as immediately useful as parallel data, should also be collected as a means for research involving alternative methods of MT development.

2.3 Existing resources

Before focusing on data collection efforts, we first investigate the language resources which were already available to us for use in training MT systems. This consists of data we collected during the Tapadóir project, as well as data made publicly available.

[9] In cases where there are multiple datasets, we use the datasets with the fewest words.

2.3.1 Tapadóir baseline data

During the Tapadóir project, we identified, gathered and processed a number of varied parallel and monolingual datasets (see Table 2.2).

Firstly, the Department of Culture, Heritage and the Gaeltacht (DCHG), as the client in the Tapadóir project, provided us with TMs generated through in-house translation. In addition, there were a number of other resources available to us. Corpas Comhthreomhar Gaeilge-Béarla (CCGB) is a bilingual dataset obtained through web crawling, and is available for download online.[10] Parallel texts from two EU bodies – the Digital Corpus of the European Parliament and Directorate General for Translation, Translation Memories – were also publicly available (these datasets are referred to collectively as EU in Table 2.2). Another dataset available to us was the Parallel English–Irish corpus of legal texts (referred to as 'Gaois' in Table 2.2). Gaois is a parallel English–Irish corpus of legal texts from the Department of Justice.[11] The language of this dataset is very technical and contains much 'legalese', or legal jargon. As well as this, we crawled 10,000 parallel sentences from the Citizens Information (CI) website,[12] referred to as CI in Table 2.2.

Corpus	**Size (GA words)**
DCHG	440,035
CCGB	113,889
EU	439,262
Gaois	1,526,498
CI	183,999
TOTAL	**2,703,683**

Table 2.2: Tapadóir baseline resources

2.3.2 Publicly available data

As well as data gathered during the Tapadóir project, we also investigated the parallel EN-GA and monolingual GA data sources available publicly (see Table 2.3).

[10] https://github.com/kscanne/ccgb

[11] https://www.gaois.ie/crp/en/data/

[12] http://www.citizensinformation.ie

This section describes sources of publicly available data that we gathered to increase the size of our training corpora.

OPUS OPUS[13] is a website for hosting publicly available language corpora for use in MT (Tiedemann, 2012a). The Directorate-General for Translation (DGT), available on OPUS, is a source of data which is updated yearly. Produced by the European Commission, it is a source of high-quality professionally translated parallel data. As highlighted earlier, due to the derogation on the production of Irish language, this data represents just a small percentage of that produced for other official EU languages. Other data from European sources are also available on the OPUS website: a parallel corpus collected from the European Constitution (EUConst) and EUBookshop. These corpora contain high-quality translations in the legal and public admin domain, respectively. The manuals for UBUNTU and GNOME are both available on OPUS. They contain bilingual text from a very technical domain, and typically contain translations of terms or short phrases containing technical jargon. Tatoeba is a parallel dataset consisting of learner-style sentences: short and containing simple grammar. The QCRI Educational Domain Corpus (QED), while consisting of translated captions from educational videos, contains similar content to Tatoeba: simple, short and clear sentences.

Name	Size (GA words)	Type
DGT	1,634,327	Parallel
EUConst	155,369	Parallel
EUBookshop	3,531,797	Parallel
UBUNTU	2,171	Parallel
GNOME	3,942	Parallel
KDE4	519,225	Parallel
Tatoeba	12,181	Parallel
QED	386,033	Parallel
Paracrawl	22,714,533	Parallel
Vicipéid	4,585,048	Monolingual GA

Table 2.3: Publicly available datasets

[13] `http://opus.nlpl.eu/`

Paracrawl Paracrawl[14] is an ongoing large-scale web crawling project designed to gather parallel data resources from the web for European languages. Although this dataset is one of the largest described in this chapter, it contains very mixed quality and can be extremely noisy at times (Defauw et al., 2019). For example, it includes crawled data from sites in which non-post-edited MT output was published. See Table 2.4, where the Irish contains non-edited MT.

Such duplicates could have a small effect on the MT output, with the duplicated translations carrying a larger 'weight' than others. If the duplicates contained noisy or incorrect segments, this could theoretically lead to errors carried through to the final output. However, as the duplicates are a tiny fraction of the overall training data collected, this is unlikely to be an issue.

English	Click gallery images for credit Featured Photo, Jonathan Brady/PA Wire Scoops/Stories Adele and Simon attend Lady Gaga's Private Concert on December 6, 2013!
Irish	Cliceáil íomhánna gailearaí do creidmheasa Réadmhaoin Grianghraf, Wire Jonathan Brady / PA Scoops / Scéalta Adele agus Simon freastal Ceolchoirm Príobháideacha Lady Gaga ar Nollaig 6, 2013!

Table 2.4: Portion of sentence from Paracrawl showing unpostedited MT output on the Irish side.

Paracrawl version 7.0 was used in the experiments described in this thesis. It should be noted that Paracrawl contains some overlap with data from citizens information, RTE and other crawled data.

Vicipéid In terms of monolingual data, data from Vicipéid[15] is both publicly available and increases in size over time. Vicipéid, belonging to the Wikimedia collection of online resources, is an online encyclopaedia written in Irish. At the time of writing, it consists of 53,034 articles published and maintained by a community of volunteers. The ethos of Vicipéid (as with other Wikimedia-related projects) is that all members of the public may edit it. While this is a positive aspect for fostering balanced unbiased articles, it can sometimes mean that the quality may

[14] https://www.paracrawl.eu/

[15] https://ga.wikipedia.org/

suffer. Vicipéid data is available to download online and provides a sizeable corpus (over 4.5 million words, see Table 2.3) of monolingual data manually produced.

2.4 Additional resources gathered

While the resources mentioned in Section 2.3.1 aided in building a promising parallel corpus, it was still necessary to build upon these datasets through further data-gathering efforts. We investigated methods of data collection to be used to supplement existing resources: **1)** directly contacting organisations which deal with Irish-language content, **2)** web crawling and (3) crowd-sourcing.

2.4.1 Public administration targeted collection

The European Language Resource Coordination (ELRC) is a European Commission-led effort to collect language resources for official EU languages, with a view to ensuring that all EU Digital Service infrastructures (such as eJustice, eProcurement etc) will be accessible in all EU languages via the eTranslation system.[16]

> "*Funded under the Connecting Europe Facility (CEF) Programme, the overall goal of the ELRC is to improve the quality, coverage and performance of the CEF Automated Translation platform in the context of current and future trans-European digital online public services.*" European Data Portal (2020).

As one of the 24 official EU languages, Irish is part of the ELRC project. Under the umbrella of the ELRC, and with the weight of an official EU project behind us, we requested language data from Irish-language organisations and public bodies that have obligations to provide Irish-language content. This involved the organisation of two workshops aimed at educating language holders on the value of language

[16] https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation_en

technology and resource sharing, as well as directly contacting or visiting organisations to aid them in identifying and sharing corpora. The two workshops were also invaluable in terms of gaining a better understanding of data management practices among public institutions.[17] In general, attendees were eager to contribute data but unsure of what exactly constitutes 'data' and had concerns surrounding licensing. Both topics had standalone sessions during the workshops, and this, as well as meeting data holders personally, was important in our data collection efforts.

The resulting datasets are extremely varied, both in terms of quantity and format. Types of data collected included staff documents, annual reports, public documents, announcements, etc. in the form of monolingual data, bilingual data and dictionaries (see examples in Table 2.5).

This collaboration provided us with a number of datasets, including data from the University Times, the student newspaper in Trinity College Dublin ('UT' in Table 2.5). Data was also contributed by Conradh na Gaeilge, an organisation which promotes the use of the Irish language ('CnaG' in Table 2.5)).

> "*The organisation runs Irish-language courses; advocates for the language rights of Irish-speakers; raises awareness about the language; hosts the international Irish-language festival Seachtain na Gaeilge le Energia; manages the Irish-language information hub PEIG.ie and the Irish-language bookshop An Siopa Leabhar; supports Raidió Rí-Rá; and much more.*" Conradh na Gaeilge (2012).

Another valuable source of data which was contributed during the course of the ELRC was Foras na Gaeilge, an Irish-language organisation which deals primarily with Irish-language support and promotion across the island of Ireland who provided us with a number of high-quality dictionaries and a large monolingual corpus (see datasets under the heading 'Foras na Gaeilge' in Table 2.5). One of the datasets contributed, the National Corpus of Ireland (NCI) is the largest corpus of monolingual

[17] See the ELRC 2017 Workshop Report for Ireland for more information: `http://www.lr-coordination.eu/sites/default/files/Ireland2/ELRC\%2B\%20Ireland\%20Workshop\%20Report-Public_0.pdf`

Irish text described in this work.

Bilingual data was also contributed by Radio Teilifís Éireann (RTÉ[18] in Table 2.5), the national broadcaster in the Republic of Ireland and Údarás na Gaeltachta[19] ('Údarás' in Table 2.5), the authority for the Gaeltacht (Irish-speaking areas of Ireland). The Irish language commissioner also contributed data via the ELRC project. "This Office, which was established in 2004, functions as an ombudsman service and compliance agency in relation to state services through Irish." – An Coimisinéir Teanga (2020).

Following on from the existing collaboration with DCHG during the Tapadóir project discussed in Section 2.1, DCHG continued to provide us with TMs created by their team of in-house translators. This data, translated by professional translators within the setting of a government department, can be described as being 'gold-standard'-, i.e. of a high enough quality that it is suitable for use as both training and testing datasets. As well as the original DCHG corpus (see Table 2.2) two additional corpora have been collected from DCHG. These are referred to as DCHG† and DCHG†† in Table 2.5.

The data collection efforts within the ELRC have been aided by the establishing of the European Language Resource Initiative (ELRI) project,[20] also funded by CEF.

> "*The main objective of ELRI is the provision of an infrastructure to help collect, prepare and share language resources that can in turn improve translation services. In particular, resources shared with the DGT will contribute to improve the EU automated translation services that are freely available to all public institutions.*" Etchegoyhen et al. (2018).

ELRI focused on the building and sharing of language resources within France, Ireland, Portugal and Spain. Within the ELRI project, French, Irish, Portuguese,

[18] `https://www.rte.ie/`

[19] https://www.udaras.ie/

[20] `https://elri.dcu.ie/en-ie/`

Spanish and English resources were collected and contributed to ELRC. ELRI also established the national relay station (NRS), which allows public administrators to upload their data on a national level, and to make use of data uploaded by other institutions in their country.[21] This data is then relayed to the ELRC-Share for use in eTranslation and, where possible, public download.

Data–set	**# of words (GA)**	**type**
DCHG†	243,372	bilingual
DCHG††	402,210	bilingual
UT	15,377	monolingual
CnaG	21,365	bilingual
Crawled	70,773	bilingual
Teagasc	32,908	bilingual
IT	57,314	monolingual
EU	483,149	bilingual
RTÉ	57,846	bilingual
Coimisinéir	129,374	bilingual
Údarás	28,395	bilingual
Foras na Gaeilge		
NCI	18,964,885	monolingual
FNGB	1,500,000	dictionary
FNGB2	549,086	dictionary
Uí Dhónaill	1,200,000	dictionary
de Bhaldraithe	1,100,000	dictionary
Téarma	180,000	dictionary

Table 2.5: Size of additional resources gathered

2.4.2 Web crawling

Web crawling is a common method for collecting bilingual data, especially for language pairs which may be lacking in resources. With both Irish and English as official languages of Ireland, many public websites have an obligation to provide online content bilingually. As such, we compiled a list of possible online sources and crawled them using the ILSP focused crawler (Papavassiliou et al., 2013), for which Irish is a supported language. The resulting corpora were often of mixed

[21] As ELRI is a continuation of ELRC, the benefits of ELRI are the same as those for ELRC: backing of an EU body to encourage more data holders to share data, possibility of onsite assistance and information sessions. In fact, the NRS has recently received funding from the Irish government to continue with data collection efforts (Gain, 2021).

quality; common issues included misalignment, comparable (similar content in each language rather than truly parallel) content, noisy data and crawling failure. In addition, while the crawler relies on consistency in webpage labelling that clearly indicates the content's language (e.g. */ga/*), we found this was not the case for many Irish websites. Therefore a pre-processing stage was necessary before adding this data to be used for MT development. A further 4,028 parallel sentences from various sources were obtained through crawling (referred to as '**Crawled**'). A list of these sites is given in Appendix A. Additional crawled datasets are '**IT**'[22] and '**Teagasc**'[23] in Table 2.5.

2.4.3 Crowd-sourcing

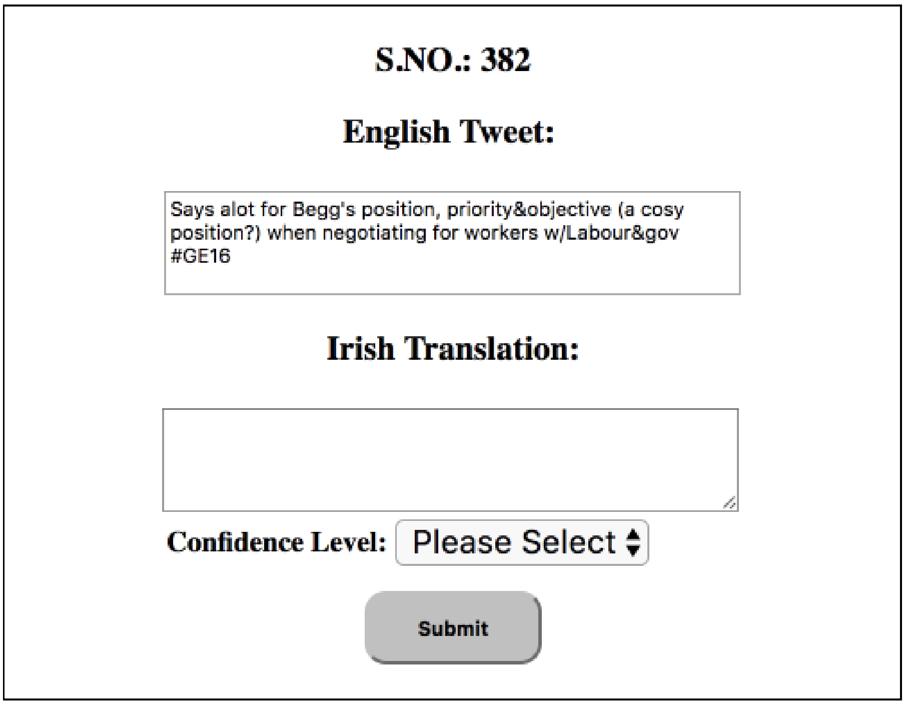
With Irish language research being low in resources – not just in terms of amount of data available but also in terms of funding and availability of skilled translators – it would be time-consuming and expensive to seek human translators to produce bilingual text for the purpose of MT. Crowd-sourcing is an alternative approach to generating new human translations, more appropriate to a low-resource setting.

We were presented with the opportunity to explore the crowd-sourcing of EN-GA translation during a collaboration on a project surrounding sentiment analysis of Irish tweets. Afli et al. (2017) used automatically- and human-translated tweets tagged with the hashtag #GE2016 to investigate the differences in sentiment between Irish and English tweets about the 2016 general election in the Republic of Ireland. We helped to facilitate the collection of human translations. Given the social characteristics of the Irish language as well as previous crowd-sourcing successes, we created an online translation interface open to the public and re-enforced its promotion with a social media campaign to elicit participant involvement from the online Irish-speaking community (Dowling et al., 2017). A screenshot of the resulting webpage can be seen in Figure 2.3.[24]

[22] The Irish Times (IT) is a national newspaper in Ireland.

[23] Teagasc is the state agency providing research, advisory and education in agriculture, horticulture, food and rural development in Ireland.

[24] Image has been slightly altered for printing purposes.



S.NO.: 382

English Tweet:

Says alot for Begg's position, priority&objective (a cosy position?) when negotiating for workers w/Labour&gov #GE16

Irish Translation:

Figure 2.3: A screenshot of the translation interface.

Through our crowd-sourcing platform, over 1000 tweet translations were collected over a period of 6 weeks (see Table 2.6). Participants were also asked to self-rate their translations by giving them a confidence rating from 1 to 10.[25] This was intended to provide some estimate of translation quality, and also to encourage more contributions. The average confidence value for both language directions can be seen in Table 2.6 and the instructions given to participants is given in Appendix B.

Language direction	Translations collected	Average confidence value
English→Irish	324	8.04
Irish→English	720	8.70

Table 2.6: Crowd-sourced translations, including average self-score rating

We also enlisted the help of a native Irish speaker to review a random portion of the tweet translations (n=180) and assign them a quality rating (1–10) based on how accurate they believed the translation to be.[26] This scoring was intended

[25] We hypothesised, through anecdotal evidence, that Irish speakers may be unwilling to participate due to a fear of not having a high enough proficiency in written Irish, even when their level of Irish is sufficient.

[26] The same scoring system as the original translator: 1 being incomprehensible, and 10 being fully acceptable in terms of fluency and adequacy.

to give us an insight into whether the crowd-sourcing participants were adequately self-scoring their translations.

Language direction	**Translations reviewed**	**Average reviewer score**
English→Irish	180	8.68
Irish→English	180	9.22

Table 2.7: Reviewer quality rating for subset of crowd-sourced data: average score for both language directions

The reviewer's average quality rating is higher (by more than 0.5) than the average rating of the translators in both language directions (see Table 2.7). Furthermore, in 71% of EN–GA translations and 82% of Irish→English translations, the reviewer deemed the translations either the same or of a higher quality than the original self-rated score (see Figure 2.4). This may indicate that participants were 'underselling' the accuracy of their translations, as was hypothesised.

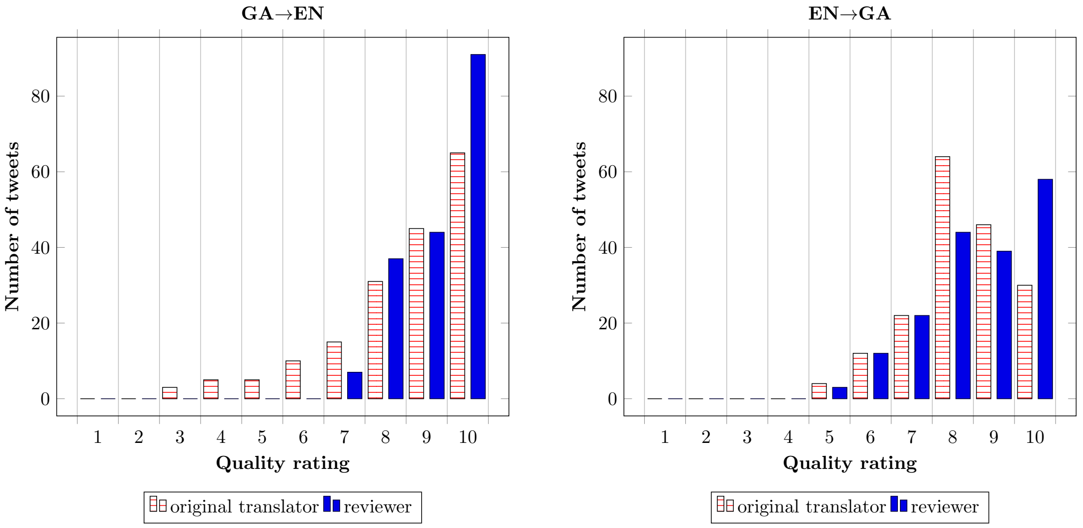


Figure 2.4: Quality ratings for Irish→English and English→Irish translations provided by the original translator and the reviewer

Findings We learned that it is possible to benefit from the altruistic nature of the Irish language community towards language cultivation, in a way that would not be possible for a majority language. It is clear that when presented with a project that has clear benefits for the Irish language, speakers will donate their time and efforts to participate. That being said, while generating awareness online is invaluable for the initial promotion of such a project, it became clear that the “hype” can die down relatively quickly if there is not a concerted effort to continue with the promotion drive. Furthermore, with both SMT and NMT relying on huge amounts of parallel data, crowd-sourcing may be too time-consuming a method for developing a sufficient corpus to train good quality MT systems.

2.5 Data pre-processing

Data gathered through web crawling efforts or direct contact with organisations is often not in a format required for use in MT training. Datasets collected via direct contact with organisations were usually in document formats such as .docx, .pdf, xlsx, among others, and data crawled from the web can see noise introduced. Therefore, much of the data needed to go through pre-processing before it could be used in MT engines. This pre-processing stage involved full cleaning (removal of formatting such as XML or HTML tags for web cleaning) and accurate alignment.

Documents were converted to plain text (.txt files) if not already in that format. For some file formats (e.g. CSV, TMX) this was a trivial task. Others required more effort. Tools such as docx2txt[27] and pdf2txt[28] were used to convert .docx and .pdf files to .txt, respectively. However, often the content of the documents was not strictly parallel, or did not follow the same pattern throughout documents. For example, a single document could contain bilingual text but switch between languages inconsistently.

Involvement with the ELRI project, discussed in Section 2.4.1, aided greatly in this process. The ELRI portal offers a pipeline of pre-processing within the NRS which could be used to convert text to aligned, cleaned parallel TMX files. A diagram of the pipeline can be seen in Figure 2.5. Contrary to open-source tools such as the ILSP crawler, the ELRI pipeline was specifically created for the processing of Irish texts in the public administration domain, and can handle various

[27] https://pypi.org/project/docx2txt/

[28] http://manpages.ubuntu.com/manpages/xenial/man1/pdf2txt.1.html

file types as input.

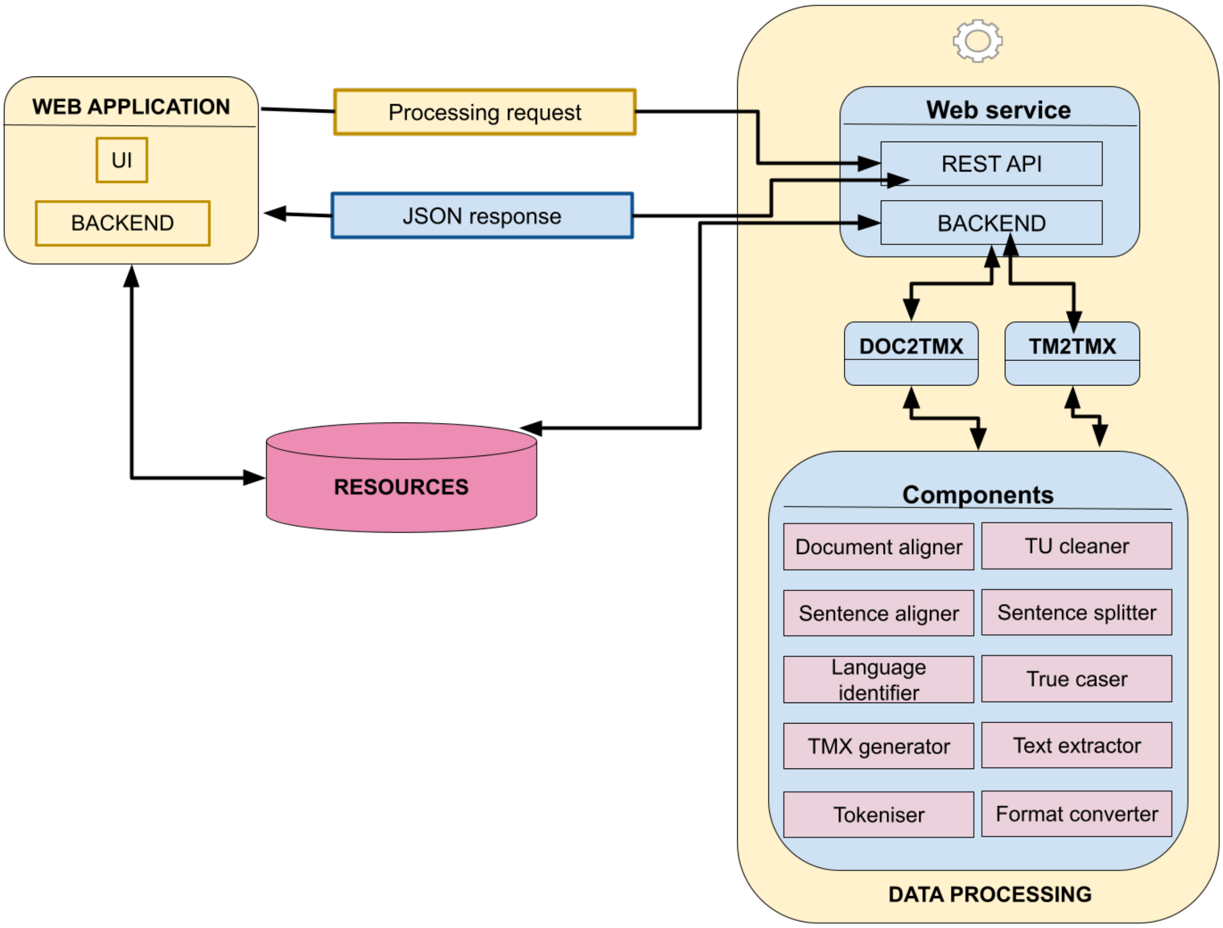


Figure 2.5: Diagram of the NRS pipeline, adapted from Etchegoyhen et al. (2018).

Anonymisation was also a task that had to be completed in the case of the DCHG data. To ensure that the data was anonymised, we replaced any email addresses, eircodes,[29] telephone numbers and PPS numbers with generic placeholder text. This step was deemed necessary by the data holders before the dataset could be contributed to the ELRC.

2.6 Licensing

When collecting language data from various sources, licensing is an element which must be taken into consideration. Although web crawling can be seen as a quick way to gather additional data, some websites contain no licensing information, or

[29] An eircode is the name for postal codes in the Republic of Ireland.

explicitly forbid web crawling.

Within the Republic of Ireland, public content is subject to an Open Data Licence (**ODL**), formerly known as a Public Sector Information licence.

> *"The European Communities (Re-Use of Public Sector Information) Regulations 2005 (as amended by SI 103/2008 and SI 525/2015) create a statutory framework for the re-use by businesses and citizens of existing information held by public sector bodies in new products and services. "Re-use", in relation to a document held by a public sector body, means the use by an individual or legal entity of the document for commercial or non-commercial purposes other than the initial purpose within the public task for which the document was produced."* [30]

This means that we could direct our web crawling efforts at organisations open under the ODL license, as well as contacting them directly, knowing that we were entitled to collect such data. This data can also be reused, and any data that we collect under ODL is being published via the ELRC.

Open source, is a category of licensing which means that the data in question may be used and reused, usually without restrictions. **Creative Commons** is one such branch of open source licences. Wikipedia, for example, is open under a **CC-BY-SA** license.

> *"CC BY-SA: This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use. If you remix, adapt, or build upon the material, you must license the modified material under identical terms"* Commons (2020).

With the help of the ELRC, some data-holders which contributed to the ELRC chose a bespoke, tailored licence to suit their individual needs. These are indicated

[30] For more information, see 'A Guide for Public Bodies on the Re-Use of Public Sector Information Regulations': `https://data.gov.ie/sites/default/files/files/PSI\%20guidance\%20for\%20PBs.pdf`

in Table 2.8 as '**Open for Reuse with Restrictions**' and '**Tailored licence**.' Such licenses allow organisations to share data in a manner that is most suited to their data type and organisation. For example, as discussed previously, DCHG data was required to undergo anonymisation before the data could be released.

2.7 Discussion and Conclusions

Name	**Size (GA words)**	**Type**	**Licence Type**
DCHG	440,035	Parallel	ODL
CCGB	113,889	Parallel	Creative Commons
EU	439,262	Parallel	Open for use with restrictions
Gaois	1,526,498	Parallel	Open for use with restrictions
CI	183,999	Parallel	Creative Commons
DGT	1,634,327	Parallel	Open For Reuse With Restrictions
EUConst	155,369	Parallel	Open For Reuse With Restrictions
EUBookshop	3,531,797	Parallel	publicly available
UBUNTU	2,171	Parallel	BSD license
GNOME	3,942	Parallel	Open source
KDE4	519,225	Parallel	Creative Commons
Tatoeba	12,181	Parallel	Creative Commons
QED	386,033	Parallel	Research only
Paracrawl	22,714,533	Parallel	Creative Commons
Vicipéid	4,585,048	Monolingual GA	Creative Commons
DCHG†	243,372	bilingual	ODL
DCHG††	402,210	bilingual	ODL
UT	15,377	monolingual	ODL
CnaG	21,365	bilingual	ODL
Crawled	70,773	bilingual	*TBD*
Teagasc	32,908	bilingual	ODL
IT	57,314	monolingual	ODL
RTÉ	57,846	bilingual	ODL
Coimisinéir	129,374	bilingual	ODL
Údarás	28,395	bilingual	ODL
	Foras na Gaeilge		
NCI	18,964,885	monolingual	*Tailored*
FNGB	1,500,000	dictionary	*Tailored*
FNGB2	549,086	dictionary	*Tailored*
Uí Dhónaill	1,200,000	dictionary	*Tailored*
de Bhaldraithe	1,100,000	dictionary	*Tailored*
Téarma	180,000	dictionary	*Tailored*

Table 2.8: All datasets collected, their word count, type and licence type.

All datasets collected, their size and the appropriate licence for each can be seen in Table 2.8. We can see that web crawling and direct contact with organisations contributed greatly in the collection of GA and EN-GA data. We recognise that Irish benefits greatly from its status as the official language of Ireland, especially from the status as an official EU language. This status has led, not just to the production of GA text as in the DGT and the DCHG, but also to the inclusion of Irish in EU-related projects such as Paracrawl, ELRC and ELRI. During our data-collection efforts we strove to take full advantage of this position, for example in targeting our web crawling efforts toward national and European public administration websites, which we know to contain high-quality data. It is worth noting that the Republic of Ireland signed the Berne Convention in 2005 (World Intellectual Property Organization, 2021), meaning that content published online without an explicit licence is automatically attributed an all rights reserved licence. With many websites failing to properly identify a licence, this could pose an obstacle for future web-crawling efforts.

In terms of advice for MT developers of other low-resource languages, we have found that direct contact with organisations can be extremely valuable where there is an organised, concerted data-collection effort. However, for language pairs without the opportunity for partnership with EU-supported initiatives such as ELRC and ELRI, the effort required (contacting organisations, acquiring licences, data conversion, etc.) could render the task difficult to complete. In this situation, we would recommend web crawling in order to gather as much data published on the web as possible, although it should be noted that it will still require pre-processing. Considering publicly-available language data, Wikipedia, a source of data available to 309 languages,[31] should also be considered as a source for the collection of monolingual data.

While crowd-sourcing can harness the positive disposition of a language community, it is most likely too time-consuming for the production of translated text. It

[31] List of Wikipedias by language: `https://en.wikipedia.org/wiki/List_of_Wikipedias`

is perhaps more useful in the production of monolingual resources (e.g. Prys and Jones (2018) and Raghallaigh et al. (2019)).

In terms of the future of Irish language corpus collection, the collection of public administration data will continue to be facilitated via the national relay station. As well as this, another CEF project which supports data collection has recently begun: PRINCIPLE.

> "*The PRINCIPLE project focuses on the identification, collection and processing of language resources (LRs) for four under-resourced European languages: Croatian, Icelandic, Irish, and Norwegian (covering both varieties: Bokmål and Nynorsk). It focuses on providing data to improve translation quality in two Digital Service Infrastructures (DSIs) – eJustice and eProcurement – via domain-specific MT engines.*" Way and Gaspari (2019, p. 112–113).

The inclusion of Irish in European CEF projects such as PRINCIPLE has shown benefits which extend beyond the timeline of the projects themselves.

> *"Since the launch of the European Language Resource Coordination (ELRC) workshops in 2016, however, some members of public administration have begun to [request the return of TMX files from an LSP], while some departments have since reported the inclusion of such a requirement in their translation contracts."* Lynn et al. (2019).

It is our understanding that these projects have also begun to establish good data management practices, which should help in the collection of Irish data in the future.

We have shown that, through harnessing the official status of Irish via EU data collection projects, web crawling, crowd-sourcing and seeking out existing resources we have succeeded in gathering a corpus of 40,364,626 GA words from public data, and added an additional 23,486,968 GA words through data-gathering efforts. In

Chapter 3 we will show how this data can be used to train EN-GA SMT and NMT systems.

Chapter 3

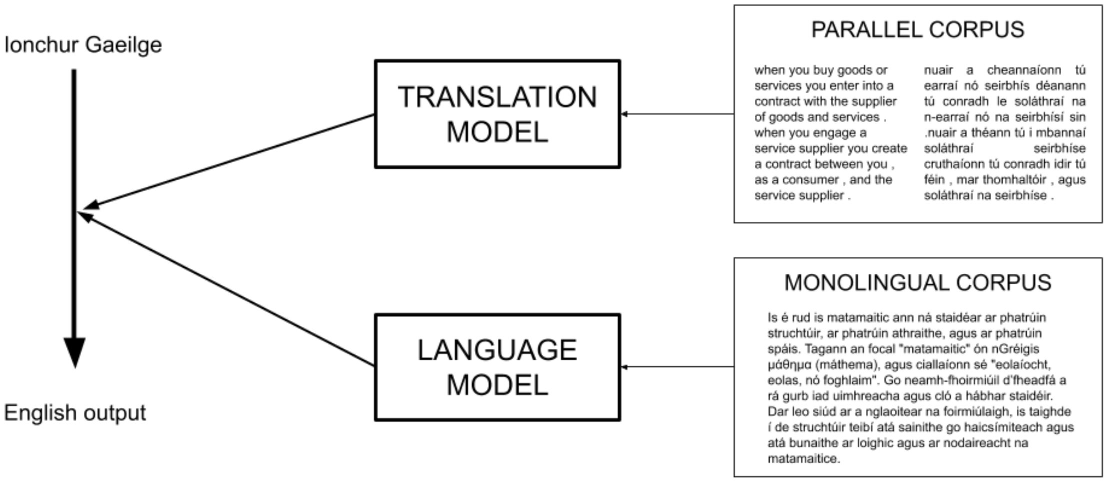
Statistical and neural machine translation

In this chapter, we provide a comparison of statistical machine translation (SMT) and neural machine translation (NMT) for English→Irish in the fixed domain of public administration via automatic metrics and preliminary linguistic analysis. We discuss the challenges for SMT and NMT of a less-resourced language such as Irish, and show that while an out-of-the-box NMT system may not fare quite as well as a tailor-made domain-specific SMT system, EN→GA NMT can surpass SMT in terms of automatic metrics with additional data and parameter-tailoring.

3.1 Introduction

Until recently, SMT (Koehn, 2009) enjoyed many years as the state-of-the-art paradigm in MT. Although challenged recently by the increasingly popular NMT, we contend that SMT can still be a valuable part of MT research. This is particularly true of lesser-resourced languages that do not have the huge amounts of data required for NMT training (Koehn and Knowles, 2017).

Building an SMT system involves training a translation model using a parallel bilingual corpus and training a language model using a monolingual corpus. Simply put, statistical methods are used to find the most probable translation of a sentence,



when you buy goods or services you enter into a contract with the supplier of goods and services . when you engage a service supplier you create a contract between you , as a consumer , and the service supplier .

nuair a cheannaíonn tú earraí nó seirbhís déanann tú conradh le soláthraí na n-earraí nó na seirbhísí sin .nuair a théann tú i mbannaí soláthraí seirbhíse cruthaíonn tú conradh idir tú féin , mar thomhaltóir , agus soláthraí na seirbhíse .

Is é rud is matamaitic ann ná staidéar ar phatrúin struchtúir, ar phatrúin athraithe, agus ar phatrúin spáis. Tagann an focal "matamaitic" ón nGréigis μάθημα (máthema), agus ciallaíonn sé "eolaíocht, eolas, nó foghlaim". Go neamh-fhoirmiúil d'fheadfá a rá gurb iad uimhreacha agus cló a hábhar staidéir. Dar leo siúd ar a nglaoitear na foirmiúlaigh, is taighde í de struchtúir teibí atá sainithe go haicsímiteach agus atá bunaithe ar loighic agus ar nodaireacht na matamaitice.

Figure 3.1: Simplified SMT diagram

based on information obtained from the translation model and the language model. This is illustrated, in a simplified way, by the diagram in Figure 3.1.

Equation (3.1) describes the original IBM word-based SMT equation, where $p(T|S)$ represents the translation model, and $p(S)$ represents the language model (Vogel et al., 1996). This equation states that $\tilde{S}$, or the best translation of the input S, is the maximum ($arg\,max$) probability of T (the candidate translation) given S. In simpler terms, the candidate translation that has the highest probability according to both the translation model and the language model will be chosen.

$$\tilde{e} = \arg\max_{S \in S^*} p(S|T) = \arg\max_{S \in S^*} p(T|S)p(S) \tag{3.1}$$

Contemporary SMT systems are no longer based on the above equation, however, with word-based systems being quickly identified as unable to produce accurate, fluent translations. Such word-based systems fall short when attempting to translate 'many-to-one' and 'one-to-many' situations, which are abundant when translating between most language pairs. For example, 'a man' in English may be translated to '*fear*' in Irish. It is more common for modern SMT systems to be phrase-based wherein phrases (i.e. combinations of words or n-grams) and not just single words are considered (Zens et al., 2002; Koehn et al., 2003). As expected, this means that the decoding process is more complex, with 2-grams, 3-grams, etc. being decoded as well as single words. As such, Och et al. (2001) propose the use of a log-linear model that allows for the integration of an arbitrary number of features. Equation (3.2) shows an updated log-linear equation based on a phrase-based SMT system as given in Hearne and Way (2011, p. 205). In this equation T represents the candidate translations and S represents the source sentence. λ_m represents a weight indicating the importance of that feature relative to the other features, and $h_m(T,S)$ is the log probability assigned to the source–candidate pair by that feature.

$$Translation = arg\,max_T \sum_{m=1}^{M} \lambda_m \cdot h_m(T,S) \tag{3.2}$$

SMT, having previously been regarded as the standard method of training MT systems, enjoys the benefits of decades of research from which to draw on, on topics such as domain adaptation, data selection and low-resource MT for example. Although SMT does rely on large datasets to build a reliable MT system, in general it is believed not to be as data-hungry as NMT (Koehn and Knowles, 2017).

> *NMT is a data-hungry approach, requiring a large amount of parallel data to reach reasonable performance.* Hoang et al. (2018, p. 21).

However, a recent study by Sennrich and Zhang (2019) has shown that a *tailored* NMT system can outperform SMT with fewer resources than previously claimed.

NMT also uses parallel text to train a translation model, but the model is trained using neural networks. There are a number of ways to train an NMT system, the most common following the 'encoder-decoder' methodology. A simplified diagram of an encoder-decoder NMT system can be seen in Figure 3.2. The input text is first encoded into a non-word representation suitable for translation, generally a vector of real numbers. This representation can then be decoded into the target-language text (i.e. translated text).

An encoder-decoder system is usually implemented using recurrent neural networks (RNNs). An RNN is a type of artificial neural network that can be thought of as a series of stacked identical networks. In an RNN, each token is fed through the

RNN until the <eos> (end-of-sentence) symbol is reached. Once the <eos> symbol has been reached, the decoding process begins. In this way, the final state of the encoder will contain information about the entire sentence. See Forcada (2017) and Way (2019) for detailed introductions to NMT for non-experts.

Some reported strengths of NMT include a perceived increase in fluency and a higher accuracy according to automatic metrics over a variety of language pairs (Bojar et al., 2016b; Castilho et al., 2017; Popović, 2017; Wang et al., 2018). Some weaknesses of NMT are a loss of semantics (the output looks fluent but has a different meaning)[1] and over-translation (the same word appearing more than once in the output) (Zhang et al., 2020).

Several studies have been conducted contrasting SMT and NMT. One such study carried out by Bentivogli et al. (2016) compared the output of SMT and NMT systems, taking English-German as a case study. Results favoured NMT over SMT, with NMT boasting a lower post-edit effort, fewer morphology errors, lexical errors and word order errors.

Castilho et al. (2017) reported on an extensive study comparing SMT and NMT, using both automatic and human evaluation methods across multiple language pairs. Results showed NMT to be promising, though not at the same level as SMT in certain areas.

> *Even though the neural model demonstrates gains in fluency, it also shows a greater number of errors of omission, addition and mistranslation.* Castilho et al. (2017, p. 118).

More recently, Mutal et al. (2019) provide a study with translators at the forefront, comparing SMT and NMT systems designed for use within the Swiss Post's language service. Their findings suggest that translators perceive there to be more errors in SMT than NMT output, but that errors in NMT output are more often disputed between translators. This supports the theory that a common error in NMT could be a loss of semantics while preserving grammatical integrity.

[1] While fluent inadequate output is a feature often attributed to NMT, it has also been shown

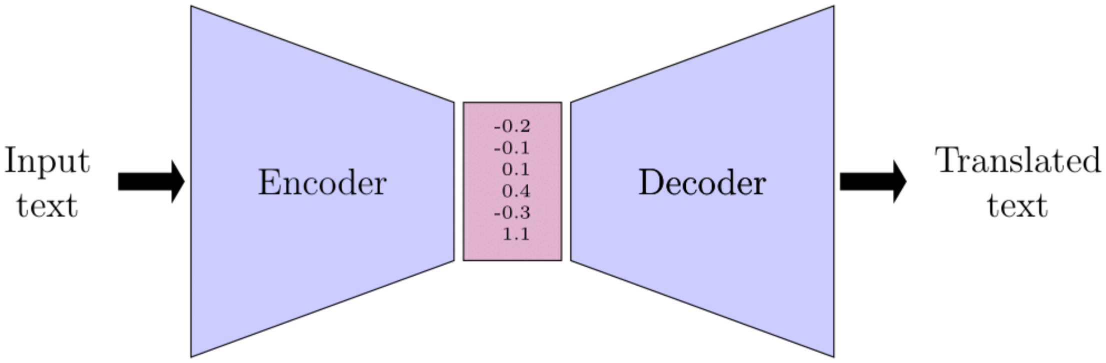


Figure 3.2: Simplified NMT diagram

In recent times, NMT has been widely hailed as a significant development in the improvement in quality of MT. However, as a technique that can be data-hungry (e.g. Jassem and Dwojak (2019) and Duh et al. (2020)), there is a concern that languages with fewer resources may not benefit to the same degree that well-resourced major languages do. As Koehn and Knowles (2017) highlight, current NMT systems can face a number of challenges when dealing with specific tasks. These challenges include low-resource languages, low-frequency words arising from inflection, long sentences, and out-of-domain texts. In order to prevent a low-resource language such as Irish being left behind in the context of these advancements, we take the first steps towards applying NMT methods to EN–GA translation.

3.2 Related Work

For many years, there have been extensive studies to show how the integration of MT within a professional translation workflow (involving post-editing by human translators, often complementary to the use of translation memory tools) improves productivity, both in industry-based and in academic-based research (e.g. Guerberof Arenas (2008) and Etchegoyhen et al. (2014)). With the introduction of NMT methods, there have been subsequent studies examining the differences between the impact that SMT and NMT have within such a setting. For example, Bentivogli et al. (2016) carried out a small-scale study on post-editing of English→German

to be present in SMT (Martindale et al., 2019) translated TED talks, and concluded that NMT had made significantly positive changes in the field. Bojar et al. (2016a) report a significant step forward using NMT instead of SMT in the automatic post-editing tasks at the Conference on Statistical Machine Translation (WMT16). More recently, Castilho et al. (2017) carried out a more extensive quantitative and qualitative comparative evaluation of PBSMT and NMT using automatic metrics and professional translators. Results were mixed overall. They varied from showing positive results for NMT in terms of improved (perceived) fluency and errors, to achieving no particular gains over SMT at document level for post-editing. While these studies were carried out on better resourced language pairs (English→German, Portuguese, Russian and Greek), they are still highly relevant in indicating the potential impact that the change in MT approaches can have in real-life translation scenarios.

Aside from examining the impact on translator productivity, there has also been increased focus on addressing the shortcomings of NMT, such as those outlined by Koehn and Knowles (2017). As such, a number of innovative approaches have emerged to this end. The application of various transfer learning methods has proven successful for certain low-resource languages (Zoph et al., 2016; Passban et al., 2017), as has the inclusion of linguistic features when addressing data sparsity that faces morphologically rich languages (Sennrich and Haddow, 2016). Luong et al. (2015) show that the use of attention-based NMT can have positive results in many aspects of MT, including the handling of long sentences.

The published study this chapter is based on is, to our knowledge, the first comparison of EN-GA SMT and NMT (Dowling et al., 2018).

3.3 Statistical machine translation parameters

In this section we outline the existing system parameters that are in place, and show the effect of different data quantities and combinations on the BLEU score of EN→GA SMT.

In previous work (Dowling et al., 2015), we experimented with different parameter settings and system tuning in order to increase the quality of the SMT output. The system settings currently in use are hierarchical reordering tables, a 6-gram language model and the introduction of an APE model.

3.3.1 Hierarchical reordering tables

Reordering table(s) are features of phrase-based translation models that inform the system how phrases should be reordered during translation (Koehn, 2009). A hierarchical reordering model for reordering tables is better able to handle larger ordering differences, by treating adjacent phrases as a single unit (Galley and Manning, 2008). These are seamlessly integrated into a standard phrase-based MT system. Previous work (Dowling et al., 2015) indicated that changing the reordering table used from phrase-based orientation to hierarchical showed a positive improvement, and so we choose to continue using this feature for current SMT systems.

3.3.2 6-gram language model

Due to the differing word orders of English and Irish (SVO versus VSO, as mentioned in Chapter 1) a common error seen in EN→GA MT is incorrect word order. This is particularly true for SMT, which only considers a fixed number of words at a time when calculating translations. This number of words is known as an n-gram. For example, if the default 3-gram is used, the MT system will consider 3 words at a time (e.g. from the example in Table 3.1: ‘Chuala an fear,’ ‘an fear leis,’ ‘fear leis an,’ ‘leis an bhféasóg’ ,’ and so on).

GA sentence:	Chuala	an	fear	leis	an	bhféasóg	scéal	nua.
EN gloss:	heard	the	man	with	the	beard	story	new.
EN sentence:	‘*The man with the beard heard a new story.*’							

Table 3.1: An Irish (GA) sentence with English (EN) gloss and corresponding sentence.

Language models based on higher order n-grams are more likely to be able to arrange short phrases in the right order than language models based on smaller *n*-grams. For instance, looking at the example in Table 3.1, a larger *n*-gram model is more likely to place the phrases '*Chuala an fear*', '*leis an bhféasóg*', and '*scéal nua*' in the correct order, assuming that the sentence has not been extracted from elsewhere in the corpus. We found that increasing the default 3-gram model to a 6-gram model aided our MT systems in this respect, and thus we have continued to employ a 6-gram language model.

3.3.3 Automated post-editing module

Usability and user experience are extremely important factors in the EN-GA MT use case. As the primary aim is to improve the speed and productivity of translators, it is crucial to produce a tool that does not hinder the user in any way. As part of the translator-developer feedback loop established with the DCHG, translators reported some repetitive errors in the MT output that were causing frustration. On closer examination, most of the errors were grammatical problems arising from Irish language morphology, most likely due to a lack of sufficient training data for the SMT system. In comparison to English, Irish has a richer morphology, such as inflected prepositions and the initial consonant mutations, which pose challenges for SMT due to data sparsity. This problem is compounded in the case of lesser-resourced languages where there are low instances of various inflected forms in the training data.

This gap in knowledge could be bridged through a number of methods such as increasing the volume of training data (where the system becomes familiar with various inflected forms of a word), byte-pair encoding (BPE, where the training data is broken into subword units before training, Sennrich et al. (2016b)) or through the introduction of post-processing module that could address simple grammatical issues on a word level basis.

To this end, we designed an automated post editing (APE) module that could address trivial spelling issues or contraction issues that challenged the SMT system.

By automatically post-editing these errors, translators can dedicate more time to more important issues such as language style. The addition of APE is intended to improve the translator user-experience and avoid any negative impact of repetitive grammatical or orthographic errors, thus creating a more enjoyable user experience.

Designing the APE module

To develop the APE module, our translator-developer feedback loop enabled us to acquire information on frequently occurring errors, and occurrences of mistranslations. On inspection, translations contained a high number of errors related to Irish language prepositions, eclipsis, lenition and contractions. This motivated the development of a set of manually written rules (in the form of regular expressions) to correct regularly occurring errors in Irish MT output. Rule sets were developed for individual prepositions and contractions and are triggered by the presence of lexical items in MT output. The APE module is split into two parts: one part which deals solely with orthographic rules, and another which addresses errors caused by grammatical case. In total there are 167 hand-written rules, which have been divided into 55 rule groups (according to preposition and error type). These scripts have been made available online.[2]

MT Errors related to orthographic rules in Irish

16 of the most common Irish simple prepositions can be inflected to mark pronominal objects (Christian Brothers, 1960; Christian Brothers, 1962), known as prepositional pronouns or pronominal prepositions. For example, it is ungrammatical in Irish for a pronominal object to occur separated from the preposition (Ó Múrchú, 2013). Such occurrences on occasion arise in the translation output, however, possibly due to a specific phrase being unseen by the MT system and subsequently translating the phrase on the individual word level. An example of an APE rule now implemented in the systems produces correctly inflected forms of these prepositions when the

[2] `https://github.com/ismisemeg/APE`

system incorrectly generates word for word translations (see Example 1).

(1) ag sinn* → againn

'with us'

Irish includes orthographical rules that aid pronunciation and reduce ambiguity from sentences, such as the rule driven by the pronunciation of neighbouring vowels. For example, if a word ending in a vowel is followed by a vowel-initial word, morphophonemic rewrite rules are applied to change the spelling to aid pronunciation (Ó Siadhail, 1989). Examples (2) and (3) show eclipsis and h-prefixing, respectively, being applied to prevent vowel elision.

(2) <u>Eclipsis</u>

(i + vowel) → (in + vowel)

i Éirinn → i**n** Éirinn

'in Ireland'

(3) <u>h-prefix</u>

(le + vowel) → (le + **h**+vowel)

le úll → le **h**úll

'with an apple'

MT Errors with Grammatical Case in Irish

The APE module is designed to correct errors which may arise due to the system's occasional incorrect choice of grammatical case. Modern Irish includes three main grammatical cases: nominative, genitive and vocative. In Irish, nouns are marked with case through various morphological changes such as lenition (e.g. *an buidéal* 'the bottle' → *dath an b**h**uidéil* 'colour of the bottle'), eclipsis (e.g. *na fir* 'the men' → *foirgneamh na **bh**fear* 'the men's building), and slenderisation or broadening of consonants (e.g. *an dochtú**ir*** 'the doctor' → *ainm an dochtúr**a*** 'the doctor's name').

The nominative form is sometimes regarded as the ‘common case’ (Christian Brothers, 1960; Christian Brothers, 1962) as it also replaces the dative and accusative cases. While the dative case is not commonly marked in Modern Irish,[3] definite nouns that are objects of prepositions still undergo an inflection process. This morphological change may also vary depending on dialect.

The Irish language has three main dialects: the Ulster dialect, Connacht dialect and Munster dialect. Inflection of definite prepositional objects (in the form of initial mutation) is realised through either lenition (Ulster dialect) or eclipsis (Connacht and Munster dialects, Ó Siadhail (1989)). From a spelling standards’ perspective, the translators in the DCHG follow the standard orthography for Irish (An Caighdeán Oifigiúil (Rannóg an Aistriucháin, 1962)), which means they should be consistent within a document, given their chosen type of initial mutation. This means that, while MT output of a lenited form of prepositional object may in fact be grammatically correct, it often requires correction to ensure consistency. Through observation of the data at hand, we chose to consistently use eclipsis as the default for the APE. If the translator wishes to instead apply lenition in a given document, they have the option to then post-edit the text manually.

In some instances, the nominal prepositional object is directly translated as a unigram (i.e. without taking into context the other elements of the prepositional phrase such as preposition and determiner) resulting in the use of an incorrectly inflected form. This is likely to be the result of the MT system backing off to translate on a unigram basis due to data sparsity in the training data. Example 4 shows the editing step required in such cases.[4] Our APE module removes the need for this correction and ensures consistency by applying rewrite rules to capture the mapping between the two dialectal forms.

(4) **MT output:** *leis an p**h**róiseas pleanála teanga*

Post-APE output: *leis an **b**próiseas pleanála teanga*

‘with the language planning policy’

[3] Few examples of the dative case still remain, e.g. *Éirinn*.

[4] Taken from actual system output.

In example (5), we show two rewrite rules, which inflect definite nouns following the prepositions *as* 'from' and *ar* 'on' to conform to the official standard spelling.[5]

(5) **(PREP + DEF. ART + NOUN) → (PREP + DEF. ART + eclipsed NOUN)**

as an baile → as an ***m****baile*

'from the town'

ar an geata → ar an ***n****geata*

'on the gate'

Rule precedence The order in which the APE rules are applied is important. We apply the orthographic rules described in Section 3.3.3 ahead of the grammatical case rules described in Section 3. Example (2) shows the steps (1 & 2) of the APE module working together on the phrase *faoin gcathaoir* 'under the chair'.

(6) **(vowel-final-PREP + DEF.ART + NOUN) → (contracted-PREP/DEF./ART + eclipsed NOUN)**

1. **Contraction**

   faoi an cathaoir → faoi**n** cathaoir

2. **Eclipsis**

   faoi**n** cathaoir → faoin **g**cathaoir

   'under the chair'

The combination of vowels in *'faoi'* and *'an'* contract to form *'faoin'* (see example 6.1). The presence of *faoin* before an ecplipsable consonant in turn triggers an initial mutation (*'****g****cathaoir'* instead of *'cathaoir'* in example 6.2). Rule precedence is clearly important here so that the orthography component of the APE module is run before the case component, resulting in the output of the first set of rules triggering the need for the second set of rules.

[5] Note that Examples 4 and 5 are similar in that they make a correction so that the eclipsed form is used, but Example 4 is showing the situation where the correct lenition, according to Ulster dialect, is changed to an eclipsed form for consistency. Example 5, however, shows an incorrect unlenited and uneclipsed form changed to an eclipsed form.

As with any language, there are exceptions to these rules. For example, in some instances, the combination of both rules can produce non-grammatical character strings (e.g. *ngC*, *mbhF*). Therefore, a small number of 'clean-up' rules were introduced to prevent the module introducing such errors. See Example (7) for a list of these rules.

(7)
1. *ngc* → *gc*
2. *ngC* → *gC*
3. *mbp* → *bp*
4. *mbP* → *bP*
5. *mbhf* → *bhf*
6. *mbhF* → *bhF*

3.3.4 Experimenting with different data combinations in English–Irish Statistical Machine Translation

The domain of the bilingual data used to train our MT system is public administration data. In keeping with this domain, the majority of the data used to train the translation model was provided by DCHG. These sources include staff notices, annual reports, website content, press releases and official correspondence. While this data may be deemed 'gold-standard', both in terms of domain and quality, we hypothesise that EN→GA SMT systems will benefit from more data, even that of different domains. Unless otherwise indicated, the GA portion of all the training data has also been used in the language model of each system.

Phase 1: Baseline datasets

With suitable baseline datasets now collected (see Chapter 2), we first experiment with different combinations of data. We establish a baseline using our original DCHG dataset, along with the CCGB corpus, crawled data and the publicly-available Gaois dataset.

We also experiment with using the GA portion of Gaois only in the language model and not the translation model (indicated with the dataset in brackets). It can be seen from Table 3.2 that a higher BLEU score is achieved when the Gaois dataset is present only in the language model.[6][7] This could be an indication that this dataset, which contains domain-specific legal text (see Chapter 2 for more details) is too out-of-domain when training MT systems for the domain of public administration.

Phase	System Training Data	BLEU
1.1	DCHG + CCGB	39.36
1.2	DCHG + CCGB + Crawled	39.20
1.3	DCHG + CCGB + Gaois	38.93
1.4	DCHG + CCGB + Gaois + Crawled	38.80*
1.5	DCHG + CCGB + (Gaois)	**39.44***
1.6	DCHG + CCGB + Crawled + (Gaois)	39.25

Table 3.2: Phase 1 training data and BLEU scores.

Phase 2: Supplementary datasets

Having secured a baseline system with promising results, in Phase 2 we experiment with supplementary datasets received from the DCHG and various European organisations (see Chapter 2), as well as the application of the APE module to MT output. The results of these experiments are shown in Table 3.3.[8] It can be seen that, although the APE module provides very minimal change in BLEU scores, the highest score BLEU score achieved is when the APE module is applied (BLEU 43.19).

However, as discussed previously, the motivation for adding an APE module is to improve the translator experience, rather than have a huge impact on the BLEU score. Generating sentence-level BLEU scores, we identified some examples where the BLEU decreased following the addition of the APE module. In Example (8) the

[6] In Table 3.2 brackets indicate that the data has been used in the language model and not in the translation model. The * symbol indicates BLEU results that are significant ($p < 0.05$) over the current baseline (DCHG + CCGB).

[7] It is worth noting that the increase in BLEU score of the highest scoring configuration is just 0.08, a difference that is unlikely to be noticed by a human translator when post-editing MT.

[8] In Table 3.3 brackets indicate that the data has been used in the language model and not in the translation model. Refer to Chapter 2 for further details on datasets used. The * symbol indicates BLEU results that are significant ($p < 0.05$) over the current baseline (DCHG† + EU).

reference translation for the phrase 'with my department's officials' is *le mo chuid oifigigh* 'with my own officials.'[9] The MT output, while matching the orthography of the reference translation (*oifigigh*, thus contributing to a higher BLEU score), is missing a h-prefix that should be triggered by the preposition *le* 'with'. The APE accurately corrects this error, resulting in an accurate and grammatical translation of the source text and removing the need for post-editing. However, the application of the APE rule lowers the BLEU score because of the increased edit distance from the reference translation. This is a clear example of how the BLEU metric can miss grammatical improvements in translation output.

Phase	**System Training Data**	**BLEU**	**BLEU+APE**
2.1	DCHG† + CCBG + EU	42.21	42.28
2.2	DCHG† + CCBG + EU + Crawled	42.24	42.33
2.3	DCHG† + CCBG + EU + Gaois	42.91	42.96
2.4	DCHG† + CCBG + EU + Gaois + Crawled	42.79	42.83
2.5	DCHG† + EU + Gaois + Crawled + (CCGB)	42.79	42.83
2.6	DCHG† + EU + (Gaois) + (CCGB)	43.11*	**43.19***
2.7	DCHG† + CCBG + EU + Crawled + (Gaois)	**43.13***	43.18*
2.8	DCHG† + CCBG + EU + (Crawled)	42.89	42.99

Table 3.3: Phase 2 training data and BLEU scores, with and without the APE module applied.

(8) *Source:* the Minister said : "I recently met with my department's **officials**.."

Irish reference: dúirt an tAire: "bhí cruinniú agam le déanaí le mo chuid **oifigigh**"

Before APE: dúirt an tAire: "chas mé le déanaí le **oifigigh** mo Roinne.."

After APE: dúirt an tAire: "chas mé le déanaí le **hoifigigh** mo Roinne.."

BLEU decrease: 25.93 to 25.68

Phase 3: Iterative adding of data

Phases 1 and 2 (see Tables 3.2 and 3.3) show a similar pattern, indicating that the Gaois dataset is of most use solely in the language model.[10] In Phase 3, we

[9] *chuid* does not trigger a h-prefix on *oifigigh*

[10] Please note that no scores statistically significant.

iteratively add additional parallel datasets to the data combination identified as the most beneficial in Phase 2, as well as another monolingual dataset, IT (Irish Times, see Chapter 2 for more details).

It can be seen from Table 3.4 that the addition of more in-domain data from the DCHG provides an increase in BLEU score, and that the APE module continues to have a small beneficial impact as well.[11] Other data sources minimally increase the score at best, and show a small decrease in BLEU score at worst.

Phase	**System Training Data**	**BLEU**	**BLEU+APE**
P3.1	P2.6 + DCHG††	44.18	44.24
P3.2	⇑ + CnaG	44.25	44.3
P3.3	⇑ + (UT)	44.25	44.31
P3.4	⇑ + Teagasc	**44.44**	**44.51**
P3.5	⇑ + (IT)	44.40	44.47
P3.6	⇑ + Crawled	44.34	44.36

Table 3.4: Phase 3 training data and BLEU scores, with and without the APE module applied.

3.3.5 Results and Analysis

Quality, when referring to MT systems, is difficult to define. Even for assessing the quality of human translation, there is no one exact equation or formula. It is quite probable that, given the same source text to translate, multiple professional translators would produce different target language translations. Therefore it follows that without an exact marker for quality in human translation, it is a difficult task to define quality for MT.

Shterionov et al. (2020) make the point that in early MT development, MT output was deemed to be of a high quality if it was 'identical to a human translation', yet nowadays, due to factors such as those discussed above, it is more fine-grained than that.

[11] In Table 3.4 brackets indicate that the data has been used in the language model and not in the translation model. Refer to Chapter 2 for further details on datasets used. The * symbol indicates BLEU results that are significant ($p < 0.05$) over the current baseline (P2.6 + DCHG††). Please note that no scores statistically significant.

O'Brien et al. (2018) argue that the notion of translation quality differs broadly between MT researchers and translation industry professionals.

> *While researchers and academics tend to focus on theoretical and pedagogic concerns related to translation quality, in most sectors of the industry translation quality assessment is broadly limited to the application of somewhat arbitrary 'one-size-fits-all' error typology models that aim to give quantitative indicators of quality.* O'Brien et al. (2018, p. 11).

Way (2013) argue that the concept of quality in terms of MT changes based on the particular use case and requirements of the particular MT system.

> *Each of the services facilitated by MT will have its own definitions of quality, dependent on the client's content and business requirements. Quality will be able to be assessed by end-users or buyers, instead of in-country reviewers.* Way (2013, p. 2).

This is an opinion shared by Lommel et al. (2014), who strove to create a comprehensive method of assessing MT (Multidimensional Quality Metrics (MQM): a framework for declaring and describing translation quality metrics).

> *In the absence of specific guidelines, "quality" can be quite quite nebulous.. something that depends on the personal expectations of reviewers.* Lommel et al. (2014, p. 457)

Therefore, we argue that a measure of quality is dependent on its use case, context and resources available. With the ultimate aim of incorporating MT output into the translation workflow of professional translators, it stands to reason that human evaluation would be best suited for evaluating quality for our use case. However, human evaluation is time-consuming, expensive and not suited to early development stages (e.g. when various parameters and data combinations may be experimented with.)

We instead chose to use automatic evaluation scores to provide a rough insight into the quality of the SMT systems as the datasets changed over time. Although the shortcomings of BLEU (Papineni et al., 2002) are well-documented (e.g. Callison-Burch et al. (2006), Post (2018), Wieting et al. (2019), and Denkowski and Lavie (2012)) it still remains the standard automatic metric in MT research. Despite its flaws, it is still very common to see BLEU scores reported in MT papers. Therefore, without the time and resources to conduct human evaluation for every data addition and parameter change, we chose to report on BLEU scores. These results are illustrated in Table 3.2, Table 3.3 and Table 3.4. In general, it is expected that with SMT, especially when dealing with low-resource languages, more data will produce higher BLEU scores. However, the results gathered in these experiments indicate that this may only be the case when the data in question pertains to a similar domain. Tables 3.2 and 3.3 show that the addition of more datasets does not always appear to have a correlation with a higher BLEU score. While one of the largest parallel datasets in our possession, the 'Gaois' dataset (see Chapter 2) contains very formal legal text, when added to the training data, it actually results in a drop in BLEU score. It is also worth reiterating that, as was shown with the addition of the APE module, a higher BLEU score may not necessarily align with a better perceived quality by human post-editors. While BLEU scores can be helpful in guiding the direction of development, if the translation experience is improved then the BLEU score is irrelevant.

These experiments have shown (i) the importance of data selection in the training of an in-domain SMT system, and (ii) that a SMT system fit for translator post-editing is achievable for EN→GA MT (Escartín and Arcedillo (2015) indicate that a BLEU score of 45+ can increase translator productivity in the case of English-Spanish MT.

Although this study shows a different target language than in our case, lacking a specific study on English-Irish MT we can take this score as a rough estimate. Future work stemming from this thesis could be to conduct a study on whether there is a correlation between translator productivity and automatic evaluation metrics for EN–GA MT.

3.4 Neural machine translation

Baseline

In order to provide a preliminary NMT baseline for EN-GA in this domain, we implement a 'vanilla' NMT system, i.e. using default parameters where possible (this system is referred to as NMT-base in Figure 3.4). We use OpenNMT[12] (Klein et al., 2017), which is an implementation of the popular NMT approach that uses an attentional encoder-decoder network (Bahdanau et al., 2014). We train a 2-layer Long Short Term Memory network (LSTM[13]) with 500 hidden neurons for 13 epochs. For the sake of comparison we use the same training data as used in the SMT system. The resulting vocabulary size is 50,002 (English) and 50,004 (Irish).[14] Note that we also apply the APE module to the output of the NMT system for comparison.

Further NMT experiments

To add to this baseline system, we also performed a few preliminary experiments to investigate the effect that altering parameters or using other methods would have on an EN-GA NMT system.

- **NMT-250** One such experiment involves experimenting with the number of hidden layers in our NMT system. We implement a smaller model, i.e. reduced the number of hidden neurons from 500 to 250. The results for this system are presented in Table 4.8 wherein this system is referred to as 'NMT-250'.[15]

[12] `https://opennmt.net/`

[13] An LSTM is a type of recurrent neural network, capable of learning long-term dependencies.

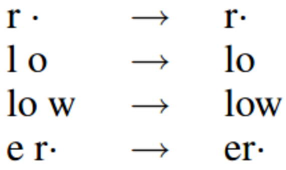
[14] These were the default, or 'vanilla' parameters used for training NMT systems with OpenNMT at the time of experimentation.

[15] In Table 4.8 the highest BLEU score and lowest TER score are highlighted in bold. The * symbol indicates BLEU results that are significant ($p < 0.05$) over the baseline (NMT-base).

- **NMT+ADAM** We also experiment with implementing the stochastic gradient algorithm with ‘Adam’. Adam is a stochastic gradient descent method which can be used for stochastic optimisation (Kingma and Ba, 2015), changing from the default stochastic gradient descent (also known as SGD) implemented in OpenNMT. This method computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients. We implement this method using the recommended learning rate for Adam (0.001) and denote this system in Table 4.8 as ‘NMT+ADAM’.

- **NMT+BPE** In order to address the inflectional nature of the Irish language, we experiment with the use of byte-pair encoding (BPE). BPE is a technique presented by Gage (1994) and adapted for NMT by Sennrich et al. (2016b). The standard 32,000 number of BPE operations is used in these experiments.

  In terms of MT, BPE is a pre-training step, where words are broken into subword units. These subword units, which are generated statistically, are not necessarily morphemes.



r ·	→	r·
l o	→	lo
lo w	→	low
e r·	→	er·

Figure 3.3: BPE merge operations learned from dictionary ‘low’, ‘lowest’, ‘newer’, ‘wider’, as in Sennrich et al. (2016b)

Figure 3.3 shows a toy example of learned BPE operations.

> At test time, we first split words into sequences of characters, then apply the learned operations to merge the characters into larger, known symbols. This is applicable to any word, and allows for open-vocabulary networks with fixed symbol vocabularies. In our example, the out of vocabuary word (OOV) ‘lower’ would be segmented into ‘low er·’. Sennrich et al. (2016b, p. 1718).

The premise of this method is that there is a higher chance of a subword being present in the training data rather than a full word, particularly if the language being translated is morphologically rich. As data sparsity is an issue especially relevant to a low-resource inflectional language such as Irish, reducing out of vocabulary (OOV) words is a promising technique. This system is referred to as 'NMT+BPE' in Figure 3.4 and Table 4.8.

3.5 Results and Preliminary Analysis

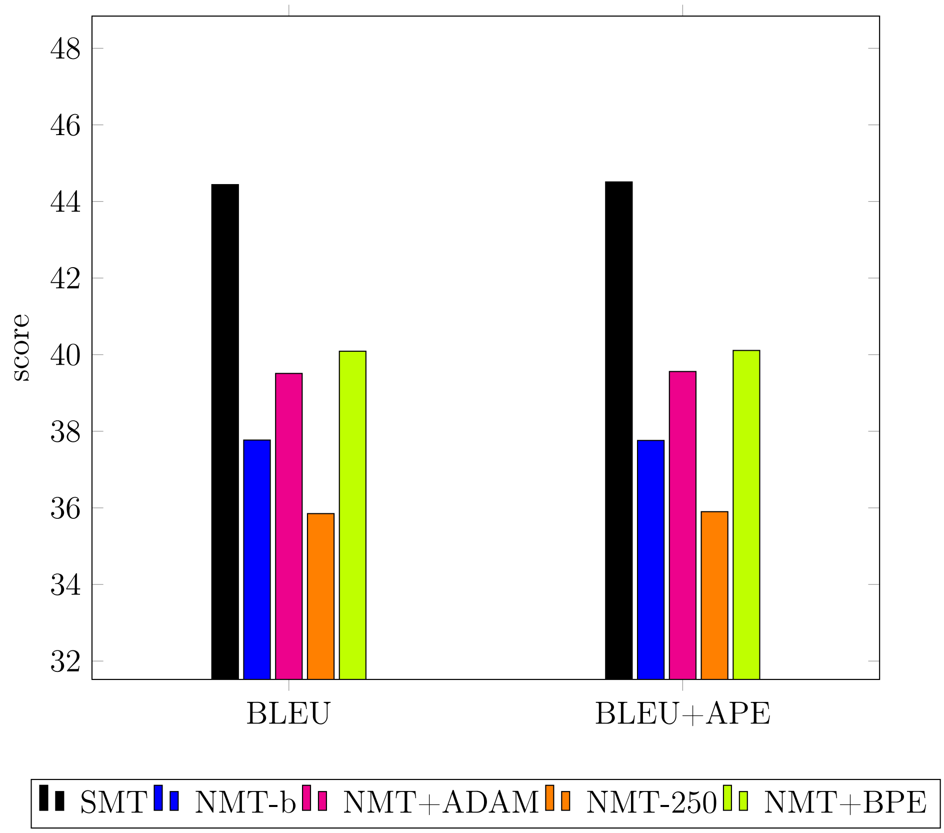


Figure 3.4: Bar graph displaying the BLEU scores of the SMT and NMT systems, with and without the APE module applied.

Both the SMT and NMT systems were tested on the same test set, consisting of 1,500 in–domain sentences randomly selected and set aside from the bilingual corpus (see Chapter 2 for further details).

Please note that following the submission of the thesis to examiners it was discovered that there contained an overlap between the test data and the training

	original BLEU	updated BLEU	Discrepancy	% reduction
NMT-base	37.77	33.7	−4.07	10.7757
NMT-250	35.85	31.9	−3.95	11.0181
NMT+BPE	40.09	33.9	−6.19	15.4403

Table 3.5: BLEU scores with corrupted test data and BLEU scores of test data with overlapping segments removed.

data, and the test data and the development data. Following this discovery, we re-estimated some of the scores to identify what kind of effect the duplicate data had on the BLEU scores. In this section, we report on the BLEU scores performed on the original test set as well as the BLEU scores performed on the new test set with overlap removed. It can be seen in Table 3.5 that the discrepancy is relatively regular. We estimate the discrepancy by removing the overlapping portions test data, which left us with a corpus of 1,120 sentences. We then calculated the BLEU scores of the new test set as well as the percentage decrease. We average the percentage decrease across all scores to give us an estimated decrease of 12.41%. Based on this, we give estimated BLEU scores in all following tables, using the average decrease of 12.41%. We thank Séamus Lankford for highlighting this discrepancy.

	BLEU	BLEU†	+APE	+APE†	TER	+APE
SMT	44.44	38.92	**44.51**	38.99	**43.31%**	43.32%
NMT-base	37.77	33.08	37.76	33.07	47.94%	47.79%
NMT+ADAM	39.51*	34.61	39.56*	34.65	46.98%	46.81%
NMT-250	35.85	31.4	35.9	31.44	50.18%	50.02%
NMT+BPE	40.09*	35.11	40.11*	35.13	46.73%	46.72%

Table 3.6: BLEU scores for SMT and NMT EN-GA systems before and after applying the automated post-editing module.

Table 3.6 shows the results of preliminary experiments for EN–GA NMT(NMT-base, NMT+ADAM, NMT-250 and NMT+BPE) in contrast to that of the best performing SMT system at the time (SMT, P3.4 in Table 3.4.)

We choose two automatic metrics to give an idea of possible changes in quality among systems: BLEU and TER. TER (translation error rate, Snover et al. (2009)) aims to measure the amount of post-editing effort required, with lower scores indicating a ‘better’ score. Both metrics are widely used to track changes in MT systems, although it should be noted that only in-depth human evaluation studies can be considered reliable sources of MT output quality.

We present our results in Table 3.6[16] and Figure 3.4. The results show that for our EN→GA use case, an out-of-the-box NMT system can establish a respectable baseline of BLEU 37.77 and TER 47.94%. However, it does not achieve the same level of quality of our tailored SMT system (showing a decrease of between 8.4 and 8.75 BLEU points – see Figure 3.4). Some alterations proved beneficial: the use of Adam as a stochastic optimisation method sees the NMT output increase in BLEU score, and the use of BPE shows an even more marked improvement. Despite these advancements, the scores are still not reaching the same quality as the SMT system.

With respect to the NMT-250 experiment, the use of 250 hidden neurons in lieu of 500 sees a decrease in BLEU score. More testing will be necessary to identify the optimal number of hidden neurons for EN-GA NMT.

We note that when the APE module is applied to the NMT output, we see very little change in BLEU score, which is in line with the trends for SMT. However, it should be noted that sentence-level analysis carried out in earlier work revealed that the BLEU score increase did not always represent better quality translation from a post-editing perspective (Dowling et al., 2016). This prompts us to carry out some investigation in this regard.

3.5.1 Sentence-level BLEU

In order to gain a preliminary insight into specific differences between EN-GA SMT and NMT, we performed a sentence-level BLEU comparison on our SMT output and NMT-base output.[17] In Examples (9)–(12), we highlight some instances where SMT outperforms NMT, and vice-versa.

[16] The SMT system corresponds to P3.4 in Table 3.4.

[17] It should be noted that sentence-level BLEU has shortcomings, even after smoothing. It is used in this context to highlight sentences with exceptional changes in BLEU scores. For a more accurate sentence-level analysis, one could use TER or the Levenshtein edit distance which do not work at a document level and therefore can be easily applied at a sentence level.

(9) ***Source:*** Islands[18]

Irish reference: na hOileáin .

SMT: na hOileáin .

NMT: Oileáin .

(NMT decrease: −69.67 BLEU)

(10) ***Source:*** when a requester agrees to amend a request that s / he has submitted, the date of receipt of the refined request is deemed to be the date of receipt of the FOI request .

Irish reference: nuair a chomhaontaíonn iarrthóir leasú a dhéanamh ar iarratas a chuir sé/sí isteach, glacfar leis gurb ionann dáta faighte an iarratais leasaithe agus dáta faighte an iarratais ar SF.

SMT: nuair a chomhaontaíonn iarrthóir leasú a dhéanamh ar iarratas a chuir sé/sí isteach, an dáta faighte an iarratais leasaithe a bheidh an dáta faighte an iarratais SF.

NMT: nuair a aontaíonn iarrthóir iarratas ar iarratas a leasú, meastar go bhfuil an t-iarratas faighte faighte ag an iarrthóir a bheith faighte.

(NMT decrease: −41.56 BLEU)

(11) ***Source:*** this also assists any possible reviews .

Irish reference: Cabhraíonn sé seo le haon athbhreithniú féideartha chomh maith.

SMT: tacaíonn aon athbhreithnithe féideartha seo freisin.

NMT: cabhraíonn sé seo freisin le haon athbhreithniú féideartha.

(NMT increase: +51.62 BLEU)

(12) ***Source:*** more about CentenaryMayo.ie :

Irish reference: tuilleadh eolais faoi CentenaryMayo.ie :

SMT: níos mó faoi CentenaryMayo.ie :

[18] This is a single word heading.

NMT: tuilleadh faoi CentenaryMayo.ie :

(NMT increase: +35.0 BLEU)

In Example (9), the SMT BLEU score is significantly higher than that of the NMT output. Delving into the translations, we can see that grammatically, NMT has correctly translated the source text (*Oileáin* 'Islands'). However, the SMT system correctly translates 'Islands' as *na hOileáin*, which literally translates as '*the* Islands'. In this domain, within the context of public administration, it is standard for 'Islands' to refer to the proper noun string '**The** Islands (of Ireland)'. This is common in place names in Irish, e.g. *An Chatalóin*, (the) Catalonia. This example highlights the value of a fixed domain, especially for low-resource MT.

Example (10) shows the translation of a longer sentence. It is clear, even to those unfamiliar with the Irish language, why the SMT output prevails in this case. The first phrase in this example is translated perfectly, when compared to the reference, meaning that it is likely that this exact phrase or very similar phrases are present in the training data, and the SMT system is therefore well-equipped to translate it. Looking at the NMT output we can see that a well-known phenomenon, not uncommon in NMT, has occurred: the translations for 'request' and 'receipt' are repeated unnecessarily (*'iarratas'* and *'faighte'*). This is sometimes referred to as 'neurobabble' or 'overtranslation' (Tu et al., 2016) and can pose problems for NMT quality.

Examples (11) and (12) show cases where NMT produces translations with a higher BLEU score than that of the SMT system. In Example (11), NMT outputs a more accurate verb (*cabhraíonn* 'assists') as opposed to the SMT output (*tacaíonn* 'supports'), and in fact achieves an almost perfect translation (*freisin* 'also' being a synonym for *chomh maith* 'as well'). It also chooses the correct inflection for *haon* 'any', which the SMT system fails to do (outputting *aon*). The *h* inflection is required following the vowel ending on the preceding preposition *le* 'with'. In Example (12), we again see NMT achieving an almost perfect translation. The translation generated by the SMT system in this case is not entirely incorrect. However, it could be argued that the NMT output is more fluent. Both of these examples highlight the strength in fluency sometimes observed with NMT.

3.6 The transformer architecture and addition of more data

Vaswani et al. (2017) propose a transformer-based approach, which focuses on attention:

> *"We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely."* Vaswani et al. (2017, p. 1).

This approach has shown promising results for low-resource NMT with other language pairs (Lakew et al., 2017; Murray et al., 2019).

Using the same data combination as in Table 3.4, we train an EN-GA system using OpenNMT, with parameters recommended by Vaswani et al. (2017). For a detailed description of the parameters used, we refer the reader to the following website for more information: `https://opennmt.net/OpenNMT-py/FAQ.html`. This configuration has been confirmed as being suitable to replicate the WMT results of Vaswani et al. (2017).

	BLEU	Updated BLEU	+APE
SMT	44.44	39.7	**44.51**
NMT-base	37.77	33.03	37.76
NMT+ADAM	39.51*	34.77	39.56*
NMT-250	35.85	31.11	35.9
NMT+BPE	40.09*	35.35	40.11*
Transformer	44.34*	39.7	44.35*

Table 3.7: BLEU scores for NMT EN-GA systems before and after applying the automated post-editing module.

The resulting BLEU scores show a marked increase over the previous best scores achieved for our EN-GA NMT systems (see Table 3.7).[19] This increase brings NMT BLEU scores even closer to that of the SMT systems. With this marked improvement in BLEU score, we continue to add data to both the SMT and NMT systems, using the Transformer architecture in the case of the latter. Rather than investigate every possible data combination, which would be too time-consuming, we instead report on systems trained with data-sets which were gradually added one by one.

Exp. number	**Data**	**BLEU**	**Updated BLEU**	**Diff over baseline**
1	*(most recent baseline)*	44.44	39.7	
2	⇑+RTÉ	44.54	39.8	⇑0.2
3	⇑+Údarás	44.76*	40.02	⇑0.42
4	⇑+Coimisinéir	44.63	39.89	⇑0.39
5	⇑+DCC	44.77*	40.03	⇑0.43
6	⇑+crawl	44.78*	40.04	⇑0.44
7	⇑+ELRI	44.58	39.84	⇑0.22
8	Exp 5 + ELRI	44.96*	40.22	⇑0.62
9	Exp 7 + Paracrawl	45.13*	40.39	⇑0.69
10	Exp 8 + Paracrawl	45.32*	40.58	⇑0.88
11	Exp 10 + EUConst	45.34*	40.6	⇑0.9
12	Exp 10 + EUBookshop	45.67*	40.93	⇑1.23
13	Exp 10 + UBUNTU + GNOME + KDE4	45.23*	40.49	⇑0.79
14	Exp 10 + Tatoeba	45.23*	40.49	⇑0.79
15	Exp 10 + QED	45.29*	40.55	⇑0.85

Table 3.8: Experiments (Exp.) showing SMT scores with addition of new data, plus the difference over the most recent baseline from Phase 3.

Exp. number	**Data**	**BLEU**	**Updated BLEU**	**Diff over baseline**
1	*(most recent baseline)*	44.34	39.6	
2	⇑+RTÉ	44.42	39.68	⇑0.08
3	⇑+Údarás	45.17*	40.43	⇑0.83
4	⇑+Coimisinéir	45.32*	40.58	⇑0.98
5	⇑+DCC	45.64*	40.9	⇑1.3
6	⇑+crawl	45.37*	40.63	⇑1.03
7	⇑+ELRI	45.86*	41.12	⇑1.52
8	Exp 5 + ELRI	46.17*	41.43	⇑1.83
9	Exp 7 + Paracrawl	46.58*	41.84	⇑2.24
10	Exp 8 + Paracrawl	46.77*	42.08	⇑2.43
11	Exp 10 + EUConst	46.82*	42.08	⇑2.48
12	Exp 10 + EUBookshop	47.0*	42.26	⇑2.66
13	Exp 10 + UBUNTU + GNOME + KDE4	47.1*	42.36	⇑2.67
14	Exp 10 + Tatoeba	47.1*	42.36	⇑2.67
15	Exp 10 + QED	47.4*	42.66	⇑2.7

Table 3.9: Experiments (Exp.) showing NMT scores with addition of new data, plus the difference over the most recent baseline from Phase 3.

It can be seen from Tables 3.8 and 3.9 that the increase in data, along with the introduction of Transformer parameters, allow the NMT EN-GA systems to

[19] In Table 3.7 the highest BLEU score and lowest TER score are highlighted in bold. The * symbol indicates BLEU results that are significant ($p < 0.05$) over the NMT baseline (NMT-base).

overtake SMT BLEU scores.[20] Whether or not this represents a true increase in quality remains to be seen, without a linguistic or human analysis of the output. However, it could be posited that a BLEU increase of more than 2 points should indicate some change in quality, presumably for the better. Other studies (e.g. Shterionov et al. (2018)) have indicated that BLEU scores may be under-reporting perceived improvements in quality for NMT systems in particular.

Interestingly, though the Paracrawl corpus is one of the largest collected for the EN-GA pair, it does not have an effect on the BLEU score of the SMT or NMT systems that one would have expected in view of its large size. This could indicate that the corpus is so noisy that any benefit gained by the inclusion of more data is being drowned out by incorrect translations introduced by noise.

3.7 Conclusion and Future Work

Our study reveals that an out-of-the-box NMT system, trained on the same EN–GA data, achieves a much lower translation quality than a tailored SMT system, at least in terms of automatic metrics. These results are not necessarily surprising given that Irish presents many of the known challenges that NMT can struggle with (e.g. long sentences and rich morphology). However, it should also be noted that automatic metrics are not necessarily congruent with actual translation quality. Shterionov et al. (2018) and Way (2019) note the shortcomings of automatic metrics, in particular when testing NMT output:

> *We show that F-measure, BLEU and TER scores do not always conform with NMT quality, as determined by human experts. Rather, they underestimate NMT quality.* Shterionov et al. (2018, p. 233).

Despite this, once the amount of training data is increased, and especially with the introduction of a transformer-based architecture, NMT indeed surpasses the

[20] In Tables 3.8 and 3.9 the * symbol indicates BLEU results that are significant ($p < 0.05$) over the current baseline.

automatic evaluation metrics of the SMT counterparts. This is inline with research which challenges the idea that NMT requires much more data than SMT to train a system with decent automatic scores (Sennrich and Zhang, 2019). This leads us to believe that EN-GA NMT can continue to progress in quality with further research and tailoring. The effects of data, with regards to the automatic metrics, appear to lessen in the case of the SMT systems, seemingly coming to a plateau (see Table 3.8).

Going forward, a more in-depth analysis of the output is needed before any difference in quality between systems can be confirmed. To this end, a human evaluation study is vital to ensure that the MT systems designed for public administration use will be optimised to enhance the task of a human translator, and will not merely be tuned to automatic metrics. This will be explored in Chapter 5.

Finally, with automatic metrics showing EN-GA SMT and NMT systems on a seemingly level playing field, investigating methods of combining systems appears to be a reasonable next step. Although data collection efforts are still underway (see Chapter 2), we can see from our experiments that more data does not necessarily mean improved automatic scores. This highlights the need to use data in creative ways. In an attempt to address both of these points, Chapter 4 will explore the use of hybrid MT to combine SMT and NMT systems and make use of monolingual data.

Chapter 4

Hybrid machine translation

In Chapter 3 we discussed our approaches to building reliable SMT and NMT systems for the EN-GA language pair and report on the effects of various parameters, dataset sizes and domains on output quality. We discussed the issue of a lack of training data and how this can contribute to data sparsity. In this chapter, we explore how monolingual data can be used to create artificial parallel data for use in training MT systems. We investigate the use of backtranslation (BT) as a method for creating artificial training data for use in EN-GA MT systems, and its effect on the quality of EN-GA MT output. We also assess whether this method could be used in the development of artificial data for a closely related language (Scottish Gaelic), and investigate the effect of this data on Irish-Scottish Gaelic (GA-GD) and English-Scottish Gaelic (EN-GD) MT systems.

4.1 Introduction

BT is a method of creating artificial parallel data through the translation of monolingual data using pre-built MT systems (Sennrich et al., 2016a). The premise of this method is that even if the data is not of human quality, the MT system can still draw benefits from the additional data. Although intuitively one might not expect that artificial data created via MT would be suitable for training MT systems, Poncelas et al. (2018) show that artificial data can be used to train MT systems both in

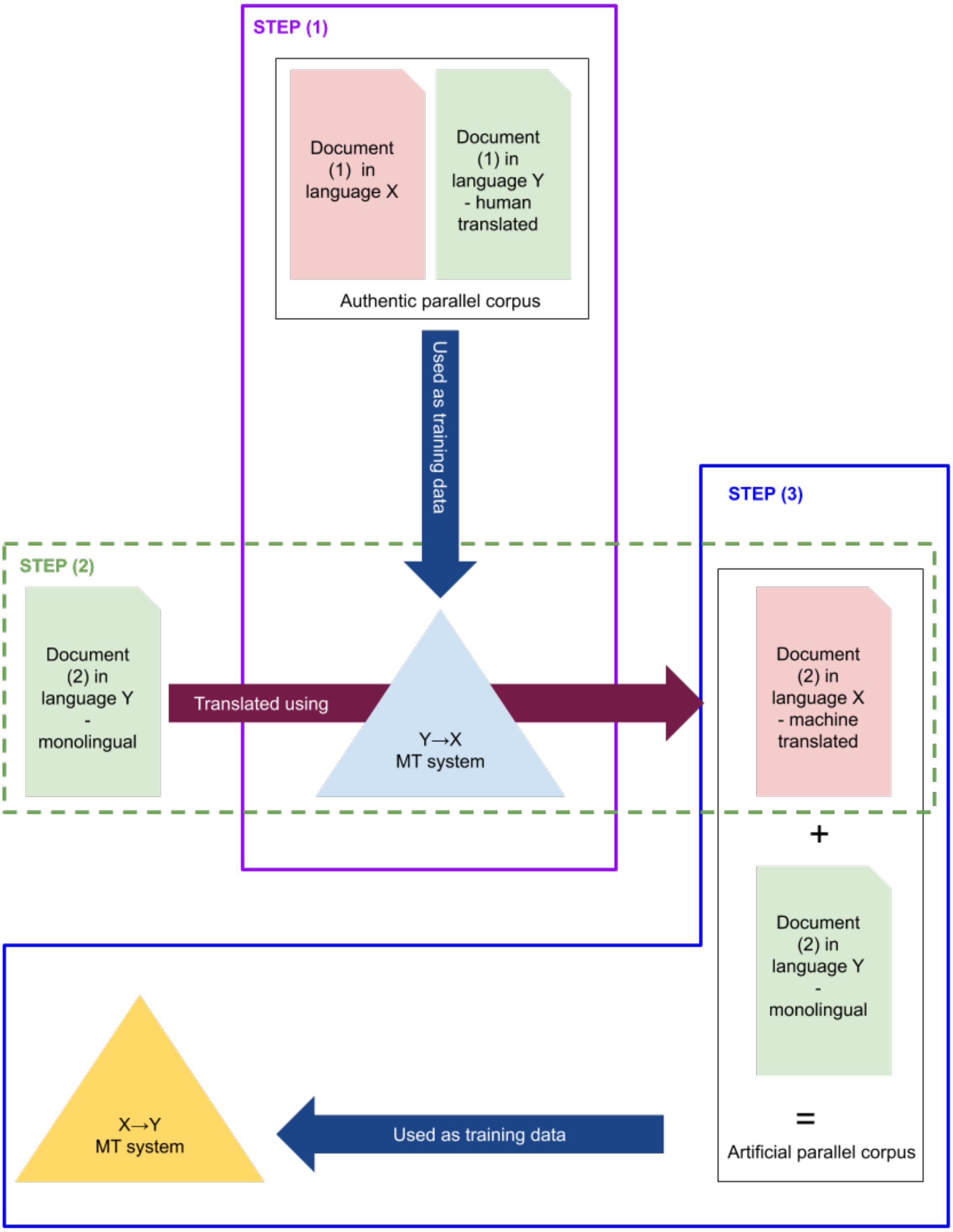

STEP (1)
Document (1) in language X
Document (1) in language Y - human translated
Authentic parallel corpus
Used as training data
STEP (2)
Document (2) in language Y - monolingual
Translated using
Y→X MT system
STEP (3)
Document (2) in language X - machine translated
+
Document (2) in language Y - monolingual
=
Artificial parallel corpus
Used as training data
X→Y MT system

Figure 4.1: Diagram of the backtranslation pipeline.

conjunction with authentic data and also with solely artificial data, and that resulting BLEU scores can exceed that of systems with only authentic data as training data. Figure 4.1 shows how backtranslation is used to convert monolingual data into artificial parallel data. In step (1), monolingual data in language X (e.g. EN) is translated to language Y (e.g. GA) using a pre-built X-Y MT system (e.g. GA-EN SMT). This creates an artificial parallel dataset. In step (2) this artificial dataset is then used to train an MT system in the opposite language direction (e.g. EN→GA NMT). (3) The resulting system can then be used to translate new documents from language Y to language X (e.g. EN-GA).

BT is a method which can be used to take steps toward addressing the challenge of MT in a low-resource scenario. While data collection efforts are extremely important in the context of resource-poor MT, it can be time-consuming and not always result in an improvement in the quality of MT output. The creation of artificial data is a quick, experimental way of increasing the amount of bilingual data available for a language pair.

The aim of the research presented in this chapter is to take steps towards addressing the area of data sparsity through backtranslation (BT).

4.2 Motivation

Despite significant efforts toward data collection for less-resourced languages (such as those discussed in Chapter 2), there is still a huge discrepancy between the amount of parallel data available in comparison to that of well-resourced languages. In the golden age of machine learning, there is a greater need for large datasets. Therefore, developers of MT systems involving less resourced languages need to implement creative solutions to get the most value from the existing data as possible.

Another advantage of BT is that it can be used to combine two different MT systems (e.g. SMT and NMT), and in theory combine the benefits of each approach, creating a hybrid system (Soto et al., 2020).

As discussed in Chapter 3, some reported strengths of NMT include a perceived increase in fluency and a higher accuracy according to automatic metrics over a variety of language pairs (Bojar et al., 2016b; Castilho et al., 2017). Some weaknesses of NMT are a loss of semantics (the output looks fluent but has a different mean-

System type	Advantages	Disadvantages
RBMT	little/no parallel data required high accuracy for related lang. pairs	requires trained linguists time-consuming & expensive
SMT	requires less parallel data than NMT can be more accurate	lack of fluency translationese common
NMT	increase in fluency can be more accurate	loss of semantics overtranslation possible

Table 4.1: A summary of some reported advantages and disadvantages of RBMT, SMT and NMT.

ing) and overtranslation (the same word appearing more than once in the output). Strengths of SMT include the need for less parallel data than NMT and sometimes a reported higher accuracy, (Castilho et al., 2017) especially in domain-specific scenarios. SMT can, however, suffer from a lack of fluency and a higher amount of 'translationese' (it is clear that the output was created by an MT system, see Toral (2019) and Vanmassenhove et al. (2019)). RBMT, although no longer widely used in production-level MT engines, can sometimes still be useful in a low-resource scenario. It requires no parallel data, just a dictionary. However, it does require sets of grammar rules which have to be hand-crafted by a computational linguist who has a high level of fluency in both languages. As a result, it can be time-consuming and expensive to create RBMT systems, which may still fall short of the quality produced by an SMT or NMT system. However, if an RBMT system is built for a language pair with similar grammar rules, it may be much more effective. A summary of some strengths and weaknesses for RBMT, SMT and NMT can be seen in Table 4.1. In this work, we aim to use backtranslation to combine SMT and NMT for EN→GA MT and combine SMT and RBMT for EN↔GD and GA↔GD.

4.3 Related Work

In this section, we discuss related work in backtranslation as well as in GD MT in general.

4.3.1 Related work in backtranslation

Koehn and Knowles (2017) highlight that current NMT systems can face a number of challenges when dealing with specific tasks. These challenges include low-resource languages, low-frequency words arising from inflection, long sentences, and out-of-domain texts.

There has been much research showing the benefits of creating artificial data for training MT systems. Despite recent increasing interest in BT for the improvement of NMT systems, the use of monolingual data as a basis for artificial parallel text is not a concept that is unique to NMT research. Although not referred to as BT, Bojar and Tamchyna (2011) describe using SMT to create artificial parallel text for use in SMT systems. A broad study involving many language pairs,[1] it reported increases in BLEU across the board, although it also noted that the positive effect was diminished as the size of the parallel dataset grew.

Sennrich et al. (2016a) present the use of BT to create artificial bilingual corpora for use in training MT systems. Using English-German as their well-resourced language pair and Turkish-English as their resource-poor language pair, they show how BT can be used to the benefit of both scenarios.

Poncelas et al. (2018) further this strand of research by assessing the effect of artificial data obtained through BT on NMT systems, when used as a standalone training corpus and also in combination with authentic parallel data.

Interestingly, the MT systems built using 1 million segments of solely backtranslated data outscored systems built on the same amount of authentic data, using METEOR as the automatic evaluation metric. This discovery highlighted BT as a promising strand of research for resource-poor machine translation.

Edunov et al. (2018) conduct a large-scale study of BT across many languages with varying amounts of resources and found that sampling methods and noise addition could be used in conjunction with BT to increase automatic BLEU scores,

[1] MT systems were built using the following language pairs: English→Czech, English→Finnish, English→German, English→Slovak, French→Czech, French→Finnish, French→German and German→Czech.

although only in resource-rich scenarios. They report that artificial parallel data alone can achieve as much as 84% of the quality of authentic parallel text within certain scenarios. The combination of sampling and BT is continued in the work of Fadaee and Monz (2018), wherein they investigate several sampling strategies for use in BT for EN→DE MT. Word frequencies and prediction losses are used to specifically target difficult-to-predict words, which is the area in which BT is most successful, according to the authors, reporting positive improvements in BLEU score.

Hoang et al. (2018) take the concept of BT and add an extra layer. They propose iterative BT, using a pipeline for creating artificial parallel data involving two stages of BT. As in standard BT, a monolingual dataset is translated using an existing MT system, and monolingual text is then translated using this MT system and used to train a new MT system. In iterative MT, this goes a step further by then using that new MT system to create more artificial data and building another MT system using the new artificial parallel corpus. Both high-resource (English-French MT) and low-resource (English-Farsi MT) scenarios are investigated, with increases in BLEU shown for both cases. They do, however, stipulate that the first MT system created by BT should be better than the pre-built MT system if it is to be used in the the 2nd step of iterative BT.

4.4 English–Irish backtranslation experiments

4.4.1 Data

The data used for building the GA-EN SMT system is listed in Table 4.2. For consistency, the same datasets were used in the creation of the baseline NMT system, with the exception of monolingual data. Monolingual data is not usually used in the training of an NMT system, yet it is required for use in BT experiments. It is preferable to use as large a corpus as possible in these experiments to maximise the amount of artificial parallel data that can be created. Datasets UT, IT and NCI contain monolingual data, while all others contain parallel data. For more details about these datasets, see Chapter 2. Being the largest corpus of monolingual Irish data, the NCI corpus was identified as a suitable starting point for the creation of artificial bilingual data.

Source	**# of words (GA)**
DCHG	440,035
DCHG†	243,372
UT	15,377
CnaG	21,365
CI	183,999
Teagasc	32,908
IT	57,314
EU	439,262
NCI	18,964,885
TOTAL	**20,398,517**

Table 4.2: Data used in backtranslation experiments

It can be seen from the final row in Table 4.2 a total of 20,398,517 GA words of data were used for MT training. Results from building engines using combinations of this data can be seen in Table 4.3.

Test data In order to test the MT systems used in these experiments, we use the same test data used in Chapters 2 – 3: a random sample of 1,500 sentence pairs held out from the DCHG portion of the training set. The test set is therefore domain-specific, and representative of the type of texts our EN-GA MT systems are most useful for (letters, reports, press releases, etc.).

4.4.2 Back Translation Experiment Set-Up

This section describes the methodology used for the creation of EN-GA artificial data through BT.

Setup and Methodology For these experiments, we follow the same methodology as illustrated in Figure 4.1 where English takes the place of language X and Irish takes the place of language Y.

Although backtranslation is more commonly carried out with an NMT system, at the time of training previous experiments Dowling et al. (2018) suggested that SMT outperforms NMT when dealing with this language pair. For this reason, we translated the NCI corpus using a GA-EN SMT system, trained with the data outlined in Table 4.2. We then train EN-GA NMT systems using differing ratios of artificial data to authentic data. This method is summarised through pseudo-code in Algorithm 1.

Algorithm 1 Pseudocode of our backtranslation methodology

```
Initialise:
$c = authentic\ corpus$
$s = size(c)$
$t = size(c)$
while $t < size(NCI)$ do
    translate $NCI$ of size $t \Rightarrow NCI'$
    $training\ data = c + NCI'$
    train EN→GA NMT system with $training\ data$
    $t = t + s$
end while
```

Applying a similar method to that used by Poncelas et al. (2018), we first begin with a 1:1 ratio of artificial versus authentic training data, and iteratively add more data until the entire monolingual corpus has been fully translated and all artificial parallel data has been added to the training data.

The GA→EN SMT system is built using Moses (Koehn et al., 2007) and configured with the best-performing parameters identified in Chapter 3. The NMT system is built using OpenNMT (Klein et al., 2017)[2] with default parameters.

4.4.3 Results and Preliminary Analysis

We evaluate each BT experiment using the test set described in Section 4.4.1, providing BLEU scores for each system. The results of these experiments are shown in Figure 4.2 and Table 4.3, in which we also provide the baseline NMT score with no artificial data added (0:1 ratio). These results show that, contrary to related

[2] https://opennmt.net/

research, the inclusion of back-translated data does not improve the BLEU score of EN-GA NMT when using these datasets and configurations. It can be seen from both Table 4.3 and Figure 4.2, that the higher the ratio of artificial to authentic data, the more the MT output decreases in BLEU score.

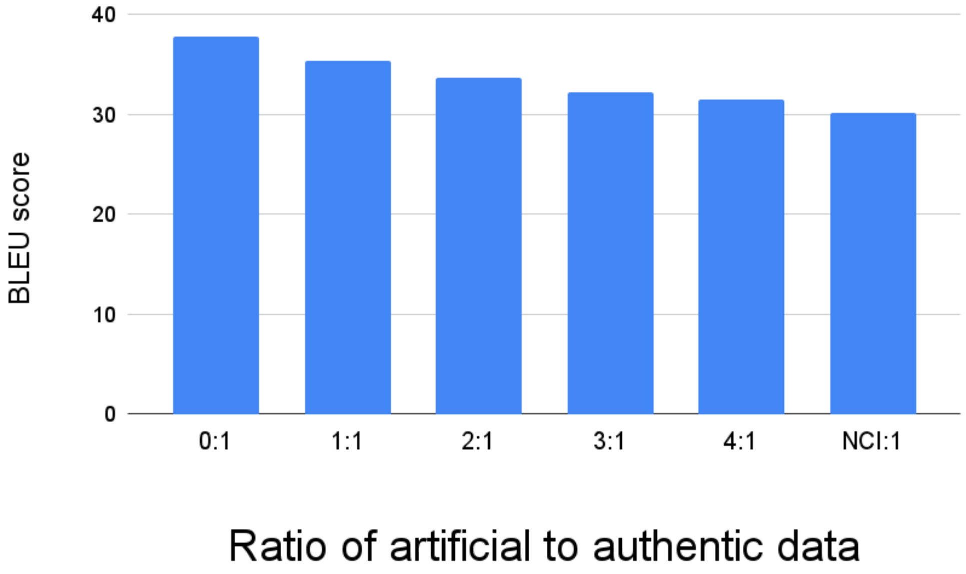


Figure 4.2: Barchart displaying BLEU scores of BT MT systems, with an NMT system without backtranslated data as a comparison.

MT SYSTEM	BLEU	BLEU†
SMT	44.44	38.92
NMT	37.77	33.08
BT 1:1	35.29	30.91
BT 2:1	33.61	29.44
BT 3:1	32.24	28.24
BT 4:1	31.46	29.44
BT NCI:1	30.18	26.43

Table 4.3: BLEU scores for BT experiments as well as earlier SMT and NMT systems.

As a marker of sufficient BLEU quality, Escartin et al. (2015) indicate that for the Spanish-English pair, a BLEU score of 45+ can increase translator productivity. Although these experiments have not been repeated with EN-GA, we can take this score as a rough guideline. The BLEU scores achieved using BT fall below this threshold, and continue to fall as more artificial data is added.

4.4.4 Sentence-level BLEU analysis

In order to gain a preliminary insight into specific changes in the MT output brought about by the introduction of backtranslated data, we performed a sentence-level BLEU analysis. This means that, rather than solely generating an overall BLEU score for the test document (as is the norm), an individual BLEU score is given for each sentence. This information can then be used to identify sentences with the biggest discrepancy in BLEU scores.

Source	**Text**	
English source	in summary, the Department's purpose is:	
Irish reference	mar achoimre, is é cuspóir na Roinne	

Artificial:authentic data ratio	**Output**	**BLEU**
0:1 (authentic data only)	mar achoimre, is é cuspóir na Roinne:	100
1:1	i achoimre, is é cuspóir na Roinne:	⇓10
2:1	in achoimre, is é cuspóir na Roinne:	⇓10
3:1	ní feidir achoimre a dhéanamh ar an méid sin	⇓90.06
4:1	ní mor achoimre a thabhairt ar an gceist seo	⇓90.06
NCI:1	ní feidir achoimre a thabhairt ar na cúinsí seo	⇓70.12

Table 4.4: A segment of MT output from MT systems with differing rations of artificial to authentic parallel training data, as well as the difference in BLEU over the system output with purely authentic training data (0:1)

In Table 4.4,[3] we see the evolution of a machine-translated sentence as more artificial parallel data is introduced to the NMT training phase. It can be seen that the baseline system with no artificial data (0:1) matches the reference exactly, and so achieves a perfect BLEU score (100). With the first introduction of artificial data (1:1), we see that the translation output changes for the worse ('i', 'in' instead of 'mar', 'because') and the BLEU score drops accordingly (⇓10). This leads to a more literal translation, which is interesting because in general it is reported that NMT is better equipped to produce fluent, rather than literal, translations (Castilho et al. (2017)). The next highest ratio of artificial data (2:1) shows a similar output, though slightly more grammatical ('i' is inflected to be 'in' before a vowel). Ratios 3:1, 4:1 and NCI:1 (just over 5:1) see the semantics of the sentence completely changed (the NCI:1 output could be roughly translated as 'a summary cannot be given for these

[3] For all examples, the ⇓ and ⇑ symbols indicate a drop or increase respectively in BLEU score over the authentic data (0:1) BLEU score

circumstances'). These examples highlight a common issue in NMT: the output looks perfectly fluent but actually displays a completely different meaning to the source text (Castilho et al. (2017)).

In contrast to these somewhat negative results, in Table 4.5 we see an example where BT has improved the NMT output, both in terms of automatic evaluation and human analysis.[4] With authentic data only, the NMT system incorrectly translates 'description' as *tuairisc*, 'report.' The first addition of artificial data (1:1) produces the MT output *cur síos* which is an exact match of the human-translated reference. This is echoed in the output of the next addition of artificial data (2:1), but changes to *cur síos ar*, 'description of' in the output of systems with ratios 3:1 and 4:1. This is another example of NMT appearing fluent (*ar* is the appropriate preposition in this situation) but containing differing semantics to the source. However, in the final addition of data (NCI:1), the MT again outputs the correct translation. This raises the question 'how much artificial data harms the MT output, and how much benefits it?' This could be an indication that if a greater amount of artificial data were added a higher level of MT accuracy could be gained.

Source	**Text**	
English source	description	
Irish reference	cur síos	
Artificial:authentic data ratio	**Output**	**BLEU**
0:1 (authentic data only)	tuairisc	30.33
1:1	cur síos	⇑ 69.67
2:1	cur síos	⇑ 69.67
3:1	cur síos ar	⇑ 90
4:1	cur síos ar	⇑ 36.34
NCI:1	cur síos	⇑ 69.67

Table 4.5: A segment of MT output from MT systems with differing rations of artificial to authentic parallel training data, as well as the difference in BLEU over the system output with purely authentic training data (0:1)

[4] It should be noted that this example contains a very short segment and is probably not representative of a typical sentence to be translated, nor of the BLEU scores of most segments. We chose this example as it was one of the few cases where the BLEU score was higher when backtranslation was applied.

4.5 Backtranslation applied to Scottish Gaelic machine translation

Although BT only showed positive results for EN-GA MT in a limited number of examples we were interested to see whether there would be more success when translating between linguistically similar languages. Scottish Gaelic (GD) is a similar but distinct language to GA, also from the Celtic language family. This section describes our efforts to leverage the greater number of language resources available to Irish to improve MT systems through BT for GD↔GA MT and build a GD-EN MT system with little or no high-quality bilingual data.

4.5.1 Scottish Gaelic background

Irish and Scottish Gaelic are Celtic languages with a number of factors in common. Both are under-resourced in terms of MT, with Irish being the better resourced. They are recognised minority languages, both in their native countries and in the EU, with English as the dominant language nationally. Although Scottish Gaelic is recognised in the UK by the Gaelic Language Act (2005),[5] neither the UK government nor the EU are legally obliged to publish Scottish Gaelic texts. This has led to a shortage in available corpora suitable for training SMT and NMT systems which leads to a major issue of data sparsity for GD MT. Without the support of laws that require the output of Scottish Gaelic content, there is the risk that GD MT will not be able to reach the same status as that of major language pairs.

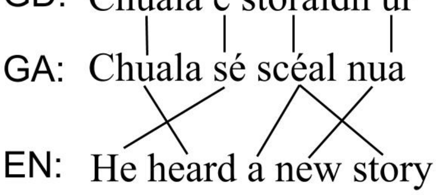


Figure 4.3: An example sentence highlighting the divergent word order between English and both Irish and Gaelic

[5] https://www.legislation.gov.uk/asp/2005/7/contents

Most linguistic differences between Irish and English would also be found between Scottish Gaelic and English. For more information we refer the reader to Dryer and Haspelmath (2013). As noted previously, translating between sentences with differing sentence structures can be a challenge for MT systems and can lead to poor quality MT output, particularly for longer sentences (Koehn and Knowles, 2017). GD, like GA, employs a verb-subject-object (VSO) sentence structure, different to the sentence-verb-object (SVO) structure more commonly seen in Indo-European languages. Figure 4.3 illustrates the similar word order of Scottish Gaelic and Irish, and how it diverges with that of English.

GD word	**English gloss**	**GA word**
creag	rock/a rock	carraig
a' c**h**reag	the rock	an c**h**arraig
creag**an**	rocks	carraig**eacha**
na cre**i**g**e**	of the rock	na carraig**e**

Table 4.6: The feminine nouns 'creag' and 'carraig' inflecting for case and number, with English glosses.

Irish and Scottish Gaelic both display richer morphology than English. The example sentence in Table 4.6 shows the inflection of the feminine nouns 'creag' (GD) and 'carraig' (GA), both meaning 'rock' or 'cliff'.[6] As discussed in Chapters 1 and 2, inflection can have an impact on data sparsity in MT.

4.5.2 Related work in Scottish Gaelic machine translation

Although it is an official language in Scotland, there has not been extensive work carried out for Scottish Gaelic (GD) MT.

There have been some efforts towards creating an RBMT system for GA↔GD using Apertium (Forcada et al., 2011), an open-source machine translation platform which uses RBMT as the underlying MT technology.[7] However, the GA↔GD Apertium module is listed as being in the *incubator* stage, which indicates that more work is needed before the MT system can be classed as being reliable.

[6] For clarity, the inflection markers (letters) in each example are displayed in bold.

[7] No research papers have been published to this end. However, Apertium is an open-source platform, and as a result all projects are available online: `https://github.com/apertium`

There has also been some previous work to create a GA→GD MT system with little or no data (Scannell, 2006). In this approach, the author built a pipeline-style MT system which uses stages of standardisation, part-of-speech tagging, word-sense disambiguation, syntactic transfer, lexical transfer and post-processing. There is also some literature surrounding the development of an SMT system for the GA–GD pair (Scannell, 2014). This approach involves training a word-based model, similar to IBM model 1 Brown et al. (1993).

In terms of NMT, research has been carried out on the GD-EN pair, in which the author uses linguistic features such as glosses to improve the system (Chen, 2018).

4.5.3 Data used in Scottish Gaelic backtranslation

In this section we describe the Scottish Gaelic and Irish language data resources used in our experiments.

Wikipedia Scottish Gaelic language Wikipedia (Uicipeid[8]) contains 14,801 articles at the time of download (04/04/2019). Pre-processing including sentence tokenising, removal of wiki-text, tags and blank lines was performed, providing us with a resulting corpus of 87,788 sentences of monolingual Scottish Gaelic. This corpus can be described as being of mixed domain, with clear, formal sentences.

OPUS OPUS (Tiedemann, 2012b) is a repository of language resources available for download from the web.[9] OPUS provides us with bilingual GA–GD and EN-GD corpora from a number of sources. Two bilingual GA–GD corpora that OPUS provides us with are the Ubuntu (655 parallel sentences) and GNOME (5,317 sentences) manuals. These are strictly within the technical domain, and often contain ‘sentences’ that are in fact 1-3 word phrases rich in technical jargon. Tatoeba, another OPUS source, is a corpus of short, simplified sentences for language learning purposes. While there was not a GD–GA Tatoeba corpus available, we downloaded

[8] https://gd.wikipedia.org

[9] http://opus.nlpl.eu/

the monolingual corpora for each language and manually aligned any matching sentences (referred to as Tatoeba-ga). OPUS also provides us with EN-GD parallel corpora from Tatoeba (Tatoeba-en), Ubuntu and GNOME.

Corpus	**# GA words**	**# GD words**	**# EN words**
Uicipeid	N/A	1,449,636	N/A
Ubuntu	20,166	25,125	N/A
GNOME	14,897	19,956	N/A
Tatoeba-ga	466	489	N/A
Tatoeba-en	N/A	2,556	2,254
EN–GA	1,433,632	N/A	1,394,726
TOTAL	**1,469,161**	**1,497,762**	**1,396,980**

Table 4.7: Number of words in bilingual (GD-EN, GD-GA, GA-EN) and monolingual (GD only) corpora used

In this work, we use the datasets used in Phase 2 in Chapter 3. This consists of 108,000 parallel sentences from sources such as the Department of Culture, Heritage and the Gaeltacht and the Citizens Information website, see 'EN-GA' in Table 4.7.

4.6 Method

In these experiments we follow the same methodology as that used in the EN-GA BT experiments described in Section 5.4 and illustrated in Figure 4.1. We carry out four sets of experiments (1, 2, 3 and 4) based on each language pair.

Experiment 1: GA→GD In these experiments (1A–C in Table 4.8), the bilingual artificial dataset is generated through BT of the GA dataset used in previous EN-GA research, as described in Section 4.4.1. The Ubuntu and GNOME data sets are used as the authentic training data.

Experiment 2: GD→GA To maintain consistency, the authentic dataset used in Exp. 1 is also used in these experiments (2A–C in Table 4.8) and the Uicipeid dataset is used as the basis of the artificial bilingual dataset (see Section 4.5.3).

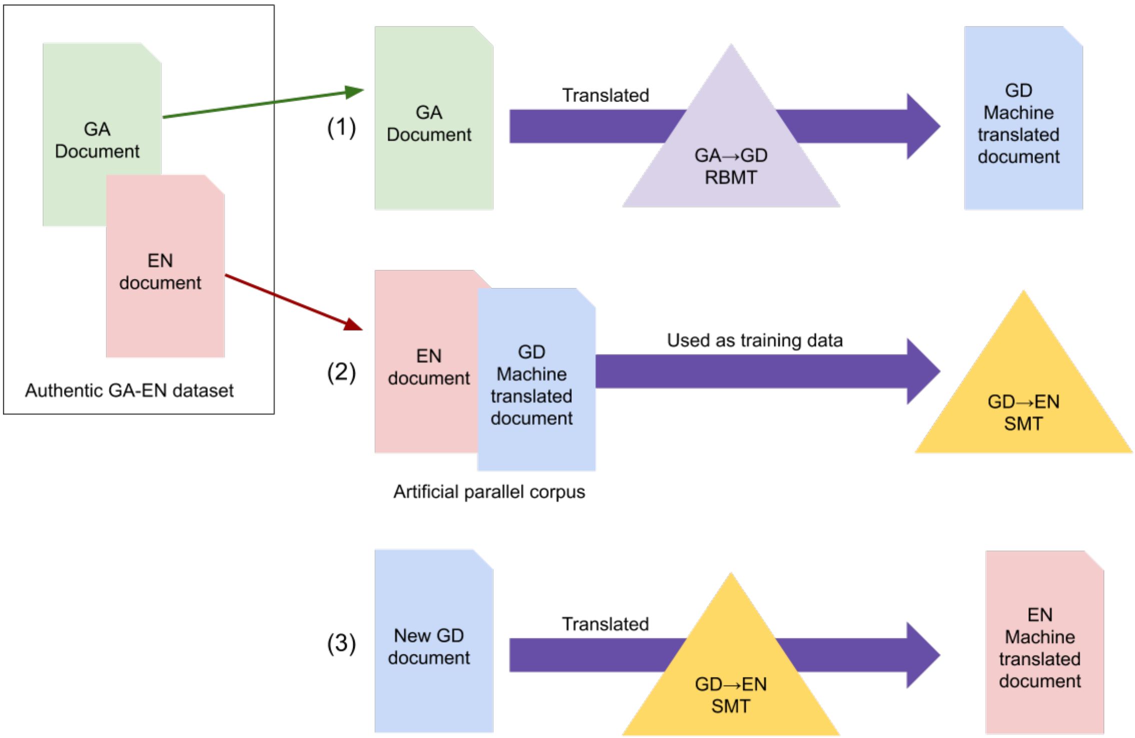


Figure 4.4: Simplified diagram of a method of using backtranslation for the development of a GD-EN MT system

Experiment 3: GD→EN With a relatively large EN-GA parallel dataset at our disposal, we chose to take this BT method a step further. In these experiments (3A–C in Table 4.8), the GA side of the EN-GA dataset is translated into Scottish Gaelic using Apertium, as in Experiment 2. However, rather than pairing the machine-translated Scottish Gaelic text with the authentic Irish text, we instead train a system using the EN portion of the authentic EN-GA dataset. This results in a GD→EN SMT system, as illustrated in Figure 4.4.

Experiment 4: EN→GD The method of generating artificial corpora is identical to that of Experiment 3, with the exception of the change in language direction. The results of these experiments are presented as experiments 4A–C in Table 4.8.

4.6.1 Building and adding to the baseline

Each experiment contains three parts (referred to as A, B and C, respectively, in Table 4.8). Part **A** involves creating a baseline by training an SMT system using only authentic data. Part **B** trains an SMT system using the artificial dataset created through BT. Finally, in part **C**, the authentic and artificial datasets are combined to train an SMT system. Systems are trained using Moses with default parameters, with the exception of the GD↔EN systems which use a 6-gram language model and hierarchical reordering tables to partly address the divergent word order between the two languages, as used in EN→GA experiments.

4.6.2 Results

We report on BLEU to provide an indication of quality for the MT systems trained. For consistency in domain, the test data for all systems comes from the Tatoeba source. It should be noted that while the source is the same, Tatoeba-ga and Tatoeba-en differ in both content and size (see Section 4.5.3).

Exp.	**Auth.**	**Artif.**	**Lang.**	**BLEU**
Apert.	N/A	N/A	GD→GA	8.67
1A	5,645	0	GD→GA	12.43
1B	0	87,788	GD→GA	16.63
1C	5,645	87,788	GD→GA	**25.45**
Apert.	N/A	N/A	GA→GD	13.73
2A	5,645	0	GD→GA	14.32
2B	0	108,000	GD→GA	17.46
2C	5,645	108,000	GD→GA	**22.55**
3A	18,785	0	GD→EN	3.73
3B	0	108,000	GD→EN	6.53
3C	18,785	108,000	GD→EN	**11.41**
4A	18,785	0	EN→GD	3.05
4B	0	108,000	EN→GD	7.03
4C	18,785	108,000	EN→GD	**10.59**

Table 4.8: BLEU scores for each experiment (Exp.), with the number of authentic (Auth.) and artificial (Artif.) sentences used to train each system. Scores are also given for the Apertium (Apert.) system used to generate the artificial data.

The results presented in Table 4.8 and Fig. 4.5 show a marked improvement in BLEU score over the baseline when backtranslated data is included as training data. We also include BLEU scores for the Apertium GA-GD module, generated through the translation of the test corpus Tatoeba-ga. Despite the low BLEU score for the Apertium GA-GD module, SMT systems trained using solely artificial data also show an increase in BLEU over the baseline. This indicates that, contrary to previous research, even if the quality of the MT system used to backtranslate is poor, it may still be possible to gain benefits from the backtranslated data. The highest automatic scores from all 4 experiment series are produced when the authentic corpus is paired with the artificial data. It is interesting to note that while BLEU scores for the EN↔GD experiments (3A-4C) are substantially lower, the same trend can still be seen. This could indicate that although the previous section did not show BT to be useful for EN-GA MT, it could still be a productive method of artificial data creation, even with linguistically different language pairs such as EN-GD.

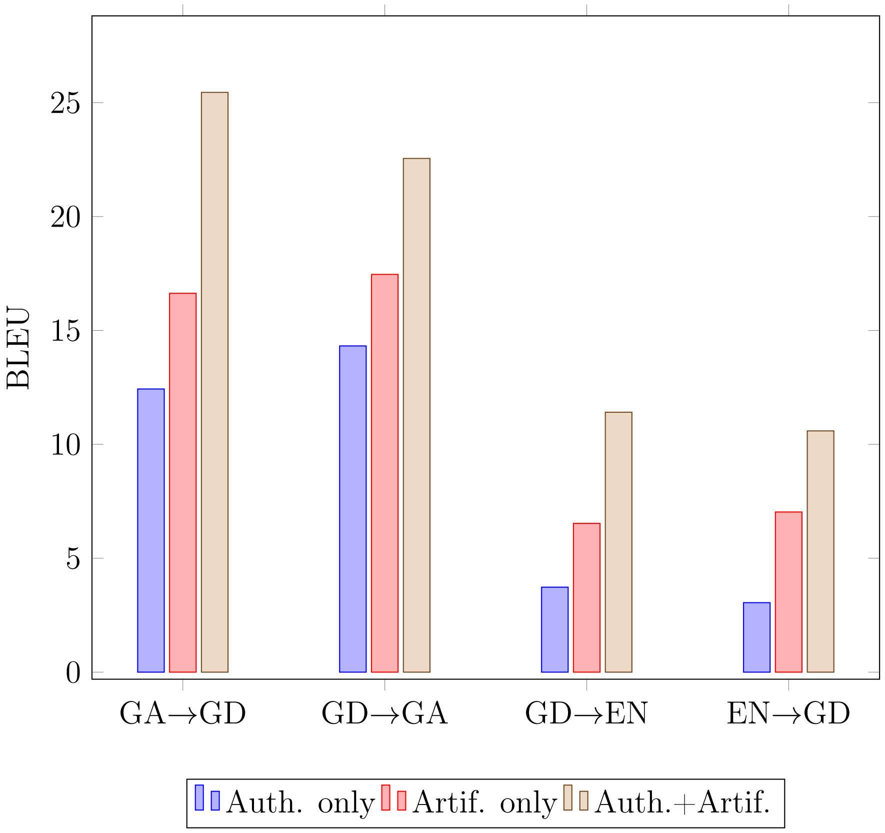


Figure 4.5: Bar chart of BLEU scores for Experiments 1 (GA→GD), 2 (GD→GA), 3 (GD→EN) and 4 (EN→GD) in Section 4.6 and Table 4.8.

4.7 Conclusions

In this chapter, we have presented preliminary results of the use of BT as a means of generating artificial data for EN→GA MT, GA↔GD and and EN↔GD MT.

Contrary to other publications on BT, we have shown from both automatic and preliminary linguistic evaluation of the MT output that BT was not successful in improving EN-GA MT using the current configuration. We can hypothesise a number of reasons for this. Firstly, perhaps our synthetic datasets were too out-of-domain, given that the NCI corpus contains a mixture of domains (i.e. literature, legal, news, etc.) and may differ too much from our domain-specific test set. Possible future work to address this issue could be to identify a monolingual dataset that is closer in domain to text from DCHG and rerun the experiments using that as a basis for the artificial parallel corpus. This could provide further insights into the importance of data selection and domain in MT.

Secondly, the original training dataset available is much smaller than those used by Poncelas et al. (2018) (one million sentences). To this end, the most obvious approach is to continue to collect parallel data through European-led projects such as ELRC and ELRI and through web-crawling efforts.

Thirdly, there may in fact be improvements in quality, but automatic evaluation metrics are not equipped to identify them. It could be seen from Table 4.5 that it is in fact possible for BT to improve some parts of the MT output. An empirical study by Shterionov et al. (2018) shows that the disconnect between BLEU and human evaluation may be as much as 50%. Way (2019) highlights the shortcomings of BLEU and conjectures that other methods of evaluation, particularly those tuned to NMT, will be necessary in the future.

We have shown that, although BT resulted in a decreased BLEU score for EN→GA MT, experiments involving GD MT saw an increase in BLEU score. It could be the case that, with just 13.73 as the highest baseline BLEU score (for a GD system trained on authentic data, see Table 4.8), there is much more room for improvement in GD MT and accordingly, data acquired through BT has a much bigger impact. It should also be considered that GA and GD are from the same language family and machine translation between these languages is expected to be an easier task than that of linguistically divergent languages. It is our hope that, with GD MT showing favourable results, this work could form a basis on which to extend to other Celtic languages and investigate whether it is useful for improving resources for similarly under-resourced languages. This would be most suitable for languages in the Goidelic language family, such as Manx Gaelic. Other Celtic languages such as Welsh and Cornish have more pronounced differences compared to Irish, though still contain similar linguistic features such as a verb-subject-object sentence structure.

With automatic metrics offering only a small insight into the quality of MT, it will be vital to use more in-depth human evaluation to gain insights into the MT output quality. Human evaluation can be used to ensure that the MT systems designed for public administration use will be optimised to enhance the task of specific human translators, and will not merely be tuned to automatic metrics. There are a variety of methods for human evaluation, not merely in the assessment of the quality of the output, but also in terms of assessing the suitability of MT systems in a post-editing environment (through calculating edit-distance, key-strokes etc., e.g. Castilho et al. (2018)). This will be the focus of the research presented in Chapter 5.

Chapter 5

Human evaluation

In previous chapters, we have shown the effects of various corpora, parameters and hybrid methods on EN-GA MT quality in terms of automatic metrics. However, while automatic metrics can be useful in gauging an idea of MT quality, they do have major limitations (Callison-Burch et al., 2006; Reiter, 2018). One limitation is that they are intended as a means for tracking the change in quality in one particular system, rather than comparing two different systems. This means that, while they can be useful for giving us an idea of differences in two systems, there may be differences that such metrics are unequipped to detect, such as those only professional translators can identify (e.g. Shterionov et al. (2018)). Another limitation is that the most widely used automatic metrics (BLEU, TER, etc.) are very static and do not allow for the dynamic and hard-to-predict nature of natural languages. For example, automatic metrics rely on one or more references to compute a score for how similar the translated text is. For low-resource languages such as Irish, it is unlikely that there is data available with multiple references (e.g. to compute multi-reference BLEU). The reference translation, therefore, is just one example of a correct translation of the source text. However, if a text was given to several professional translators, it is extremely likely that they would return differing outputs. This means that it is plausible for MT output to be a correct translation, but not match the one given reference. Therefore if automatic metrics are falling short, it is important for human evaluation to be conducted to achieve a clearer view of both the quality of EN-GA MT in general and the differences of quality between EN-GA SMT and NMT. As there have been no human evaluation studies of EN-GA MT reported to date, the main aim of the research presented in this chapter is to provide the first human evaluation study of EN-GA MT.

In contrast to other language pairs (e.g. EN-ES, FR-DE, etc.), EN-GA MT, with the majority of Irish speakers holding native-level quality of English,[1] holds the most value as a tool for aiding professional translators rather than as a gisting tool.[2] Accordingly, the study presented in this thesis aims to recreate the characteristics of the environment in which EN-GA MT is most useful – as a tool to be used by professional translators in the domain of public administration – in order to assess the usefulness of EN-GA MT in this context.

5.1 Overview of Human Evaluation

Human evaluation, in terms of MT, is when a human assesses the quality of MT output. This assessment can take many forms, e.g. ranking of MT output, annotation of incorrect parts of speech or post-editing of MT output.

Human evaluation has long been a central part of MT research, gaining particular recognition in recent years, especially when drawbacks of automatic metrics in evaluating NMT have been highlighted (e.g. Poncelas et al. (2018)). However, human evaluation is (1) more expensive and (2) more time-consuming than automatic evaluation and can therefore be overlooked in MT research.

One reason that human evaluation is essential in MT research is that MT is generally built to be used by humans – either by the general public as a gisting tool or by professional translators as a translation tool. In the second scenario, professional translators can choose to post-edit MT output rather than translating

[1] See this report from the Irish 2016 census for more information: https://www.cso.ie/en/csolatestnews/presspages/2017/census2016profile10educationskillsandtheirishlanguage/

[2] An example of MT as a gisting tool is when MT is used online to get an idea of what the text means in the target language. The user knows that the output will not be perfect but it is cheaper and more realistic than hiring a professional translator and is not intended to be published in the target language.

from scratch. As mentioned in the previous section, we argue that EN→GA MT is most useful for aiding professional translators. Therefore, it follows that these same professional translators, rather than automatically-gained scores, are best placed to inform researchers on the quality of EN→GA MT .

Domain is an important factor in MT PE. Moorkens et al. (2018) report on English to Catalan translators' perceptions of literary PE, comparing translation from scratch, editing NMT output and editing SMT output. They found that all translators preferred to translate from scratch, despite producing faster translations when post editing NMT citing a restriction on creativity among their reasons. Studies such as this highlight that although temporal effort is a robust indicator of productivity it does not take into account translator preference. While translator preference is not an easily measurable metric that can be shown to increase productivity, we believe that it is a factor that language service providers should examine. There is still a huge lack of skilled EN-GA translators (Lynn et al., 2019) and we argue that translator job satisfaction is an important consideration when training MT systems to be used by human translators. As in previous chapters, the experiments in this chapter focus on the domain of public administration. Although this domain is more fixed, with less room for creativity, we still believe that it is important that translator satisfaction be at the forefront.

In a study on MT adoption among DGT translators, Cadwell et al. (2016) posit that ergonomic factors play a role when translators chose whether to post-edit MT output or translate from scratch. This study, which involved translators of all official EU languages, argued that translators' needs and limitations should be considered when offering MT and that all translators will have differing experiences with MT.

Many methods of carrying out human evaluation studies have been examined, which can be defined as either judgement-based or measurement-based techniques. Judgement-based methods such as the use of Likert scales and ranking rely on the judgement of the human evaluation participants, usually translators or native speakers of the target language (Castilho et al., 2018). These methods can be useful as they are often quick and can easily provide evaluations of multiple MT outputs at a time. However, judgement-based methods can suffer due to subjectivity, and evaluation is usually on a surface level. Measurement-based techniques for human evaluation include annotation and post-editing (Bentivogli et al., 2018). Some benefits of measurement-based techniques may be a more fine-grained and less subjective analysis. However, these methods can be time-consuming and require participants with a high skill level in the target language, presumably translators who have experience post-editing MT output. Therefore these methods can be expensive.

Among judgement based evaluation techniques, limited survey approaches are common.

"A limited survey approach is frequently adopted in machine translation quality assessment where the concepts of adequacy and fluency have been subject to Likert-type scales for some time now." – Saldanha and O'Brien (2014), p. 104.

However, as mentioned earlier, judgement-based methods such as Likert scales can open the study up to issues of subjectivity and (as it is usually the case that each individual sentence is assessed) can be very time-consuming (Saldanha and O'Brien, 2014). In a human evaluation study focused on evaluating fluency and adequacy, Koehn and Monz (2006) found that participants in PE studies may have difficulties assigning numerical values to MT output even if explicit guidelines are given and that long sentences are particularly difficult to assess.

Ranking is another method commonly used in MT human evaluation studies. Ranking is when candidate translations from two or more MT systems are presented to participants, who are then asked to order them in terms of adequacy, fluency, overall quality, etc. This method is less time-consuming and can provide immediate feedback on which system has the better quality. However, the problem of subjectivity remains, especially when the differences among systems are not pronounced.

"Subjectivity is also an issue with ranking systems, especially if the criteria expect evaluators to estimate how much effort might be involved in revising or post-editing each translation." – (Saldanha and O'Brien, 2014), p.102.

Direct assessment (Graham et al., 2013) is another method of judgement-based human evaluation and is used in WMT shared tasks. It involves using only the MT output to perform the evaluation. One benefit of this method is that participants, therefore, need only be proficient in the target language and are not required to be bilingual.

"Translation adequacy is structured as a monolingual assessment of similarity of meaning where the target language reference translation and the MT output are displayed to the human assessor. Human assessors rate a given translation by how adequately it expresses the meaning of the reference translation on an analogue scale corresponding to an underlying 0-100 rating scale." – (Bojar et al., 2016c).

Direct assessment can also be used to assess fluency in a similar manner to those mentioned above, though the reference translation is not given and participants are instead asked how fluent the output is.

Measurement-based approaches are also common in human evaluation. An example of a measurement-based approach in human evaluation of MT systems is to annotate errors at a phrasal or token level. This method requires detailed instructions and trained linguists. This method provides the most fine-grained analysis of MT quality, although it is very time-consuming and, accordingly, very expensive. Results can be used to estimate the percentage of output that contains errors and which category of errors are most prevalent. The Multidimensional Quality Metric (MQM) Framework (Lommel et al., 2014) is one such method of analysing MT output and can be used to provide a numerical value (MQM score) for the output.

A human evaluation study can also be carried out by providing the translators with MT output and asking them to post-edit it. One benefit of this method is that subjectivity can be decreased; data gathered through translator post-editing (e.g. time spent per segment, number of keystrokes, etc.) is used to assess the MT system rather than the participant being required to give a judgement per word/segment. It is also faster than error annotation and requires less training, particularly if the translators already have experience of post-editing MT output. It is also the method which is closest to the situation in which MT is intended to be used, and as a result translators' opinions (of output quality, fluency, etc.) can also be elicited. One of the aims of the study outlined in this chapter is to investigate the measurable usefulness of EN→GA in a professional translation capacity. For these reasons, we see post-editing as the human evaluation method that suits the needs and intended outputs of this study.

5.2 Motivation

Human evaluation is a vital component of MT research (Castilho et al., 2018), with many of the major MT conferences including a translator track to encourage such publications. They are especially valuable in low-resource or minority contexts (e.g. Spanish-Galician MT (Bayón and Sánchez-Gijón, 2019), Russian-Japanese MT (Imankulova et al., 2019)) where the language pairs may be overlooked by global MT companies.

While there has been previous research on improving EN→GA MT (see Chapter 3, as well as Arcan et al. (2016) and Defauw et al. (2019)) to date there have been no publications describing a human evaluation study for EN→GA MT.

Despite MT having been established as a useful tool in the workflow of a professional translator, it is not yet the norm for Irish translators, whether freelance or within a translation company. A recent study by Moorkens (2020), p.66, reported that "...few participants appear to use MT at present..." It may be the case that MT is not an area that is being invested in, or that EN-GA MT is not available as a tool for the majority of professional translators. Without MT PE as the norm in professional EN-GA translation, it is also possible that translators are wary when it comes to EN-GA MT and lack experience post editing MT output. In terms of Irish translators' attitudes to MT, Moorkens' extensive survey reports varying attitudes between translators based on terms of employment, with freelance translators appearing to be poorly disposed towards MT.

There have been comparisons of SMT and NMT since NMT first emerged in the field. The conference on machine translation (WMT) regularly features both systems, with human evaluation at the forefront (Bojar et al., 2016b; Ondřej et al., 2017; Barrault et al., 2019). Castilho et al. (2017) describe an extensive comparison of SMT and NMT using both automatic metrics and human evaluation. Mixed results overall highlight the need for language-specific human evaluation studies. In Chapter 3, a preliminary comparison of EN→GA SMT and NMT (Dowling et al., 2018) showed that SMT fared better than NMT in terms of automatic metrics. As discussed in Chapter 3, more recent publications (Defauw et al., 2019; Dowling et al., 2019a) show a more positive picture for EN→GA NMT, but without a direct comparison to SMT. As well as providing the first reported human evaluation study of GA MT, another aim of this study is to provide a human-derived comparison of EN→GA SMT and NMT. The SMT/NMT comparison presented in this chapter will take into account both the quantitative metadata gathered during the study (time per segment, number of keystrokes, etc.) as well as the qualitative opinions and recommendations of the participants obtained via a survey.

5.3 MT systems set-up

To compare SMT and NMT through human evaluation it is first necessary to train a system of each type using the same training data. This section describes the data used in building both MT systems, their specific parameters and the automatic evaluation scores generated for each.

Table 5.1 shows the sources and number of GA words of all datasets used to build the SMT and NMT systems used in this study.[3] For more details on these datasets, see Chapter 2.

[3] This data combination relates to experiment 8 in Table 3.9 as this was the highest scoring data combination at the time of carrying out the human evaluation study.

Source	# words (GA)
DCHG	1,085,617
EU	439,262
CI	183,999
Crawled	70,773
CnaG	21,365
Teagasc	32,908
UT	15,377
IT	57,314
Paracrawl	22,714,533
ELRC	415,648
ELRI	628,669
TOTAL	**23,754,020**

Table 5.1: Source and number of Irish words of data sources used to build the MT systems described in this chapter.

5.3.1 Test data

In keeping with Chapters 3 and 4, we use the same test data as in other EN-GA MT experiments. 1,500 sentences of gold-standard data,[4] with an average sentence length of 20 words per sentence, were held out from training data in order to perform automatic evaluation. This data contains extracts from DCHG sources such as official correspondence, public announcements, website content, etc.

5.3.2 Statistical machine translation parameters

When training the SMT system, we use the best-performing parameters identified in Chapter 3. Moses (Koehn et al., 2007), the standard tool for building SMT systems, along with the data described in Section 5.3, is used to train our SMT model. KenLM (Heafield, 2011) is used to train a 6-gram language model using the GA portion of the parallel data, as well as the monolingual GA data. This wider-context language model (3-gram is the default) along with hierarchical reordering tables are used in an attempt to address the divergent word orders of EN and GA (EN having subject-verb-object and GA having verb-subject object word order.) See Chapter 3 for a more detailed description.

[4] Professionally translated data within the same domain (from the DCHG corpus).

5.3.3 Neural machine translation parameters

As in Chapter 3, as well as other research on EN-GA NMT (Defauw et al., 2019; Dowling et al., 2018), we use OpenNMT (Klein et al., 2017) as the basis for training our NMT system. We choose to implement a transformer-based approach (Vaswani et al., 2017), which has shown promising results for low-resource NMT with other language pairs (Lakew et al., 2017; Murray et al., 2019). We use parameters recommended by Vaswani et al. (2017).

5.3.4 Automatic evaluation

Automatic evaluation metrics, while best used to track developmental changes in one particular MT system over time, can also be used in an attempt to gauge differences in quality between two different MT systems. Coughlin (2003) showed that in certain circumstances, human assessments can closely correlate with automatic metrics. In this study we generate scores using the following automatic metrics: BLEU (Papineni et al., 2002), TER (Snover et al., 2009), CharacTER (Wang et al., 2016) and ChrF scores (Popović, 2015).

	BLEU↑	BLEU↑ †	TER↓	ChrF↑	CharacTER↓
SMT	45.13	39.53	43.51%	66.26	0.29%
NMT	**46.58**	40.8	**40.85%**	**67.21**	**0.28%**

Table 5.2: Automatic evaluation scores for the SMT and NMT systems used to generate MT output, rounded to 2 decimal places.

With automatic evaluation, the source side of the test data (EN) is translated using the MT system. BLEU and TER both compute scores by comparing words in the MT output to those in the GA portion of the test data. CharacTER and chrF, however, compute a score based on a character-level comparison, which can be more accurate for inflected languages.[5]

All these methods work by computing the score for each segment, and then averaging the score across segments for the entire document. This can have limitations

[5] Please note that no scores statistically significant.

when segments vary in length, e.g. one segment could be a single word as a heading and another could be a full sentence.

Table 5.2 shows the BLEU, TER, CharacTER and ChrF scores for the SMT and NMT systems used in this study.[6] These scores can then be compared to the results provided through human evaluation. Both BLEU and ChrF are precision based, with higher scores indicating higher precision and, in theory, higher quality. This is indicated with a ↑ in Table 5.2. TER (translation error rate) and CharacTER (TER on character level) are error-based metrics. Accordingly, a lower score represents a lower error rate, indicated with a ↓ in Table 5.2.

It can be seen from Table 5.2 that the NMT system achieves a better score across all four metrics, whether calculated on a word or character level.

5.4 Study methodology and set-up

As discussed in Section 5.1, one of the aims of this study is to provide a comparison of EN→GA SMT and NMT with similar circumstances as its intended use (as a PE tool for the creation of public content). This section describes the set-up and methodology of the PE task and related survey.

5.4.1 Post-editing tool and guidelines

Post-editing tool (PET) (Aziz et al., 2012) was chosen as the software with which to collect data for this study. There are some limitations to the use of PET. For example, it does not model exactly the usual work environment for a post-editor and it does not collect mouse moves, only keystrokes. Some advantages of PET are that it is freely available online and specifically designed for use in human evaluation studies surrounding MT. For these reasons, we choose PET as the tool with which to carry out this preliminary study. We configure PET with the default parameters and compose guidelines and instructions for the participants. For example, participants

[6] In Table 5.2 the best score in each column is highlighted in bold.

were permitted to use dictionaries while PE the output, but were not permitted to use another MT tool. The guidelines were written in Irish, the target language of this study, available in Appendix C. For accessibility to non-Irish speakers interested in this study, these were translated into English and are attached in Appendix D.

5.4.2 Pilot study

Prior to the main study, we conducted a pilot study to ensure that the tool was set up correctly and to test the robustness of the guidelines. Two Irish linguists each post-edited 10 machine-translated sentences. We then updated the guidelines as per the feedback of both pilot study participants.

5.4.3 Data Splits

Two subsets were extracted from the test data described in Section 5.3, each containing 100 EN sentences, and then translated with the SMT and NMT systems described in 5.3.2 and 5.3.3 respectively. With the merits of document-level translation raised in recent MT research (Werlen et al., 2018; Toral et al., 2018) and the importance of context in work using MT for dissemination, we choose to keep the sequence of sentences, rather than extract each of the 200 sentences individually at random. We refer to each of these subsets as ‘Job 1’ and ‘Job 2’ respectively.

To investigate the differences between SMT and NMT, we calculate standardised type-token ratio (STTR) with the outputs.[7] Table 5.3 shows that, although a small difference can be seen between jobs for both systems, on average both MT systems have a very similar STTR.

System	**Job 1**	**Job 2**	**Average**
SMT	41.71	42.69	42.20
NMT	43.84	41.33	42.59

Table 5.3: Comparison of STTR between SMT and NMT outputs normalised per 1000 words

[7] Type-token ratio normalised per 1,000 words.

5.4.4 Participants

Some human evaluation studies may use crowd-sourcing platforms (e.g. Figure Eight,[8] Mechanical Turk[9]) to reduce costs, although if used there is a possibility that quality may be affected. With a high demand for EN-GA professional translators, it is unlikely that experienced translators would participate in such platforms. Moorkens and O'Brien (2015) explored the trade-off between using novice and professional translators in PE studies and reported that although novice translators were more enthusiastic and willing to engage in research, professional translators produced more efficient work.

Moreover, as mentioned earlier, EN-GA MT is most likely to be used as a tool to help publish translated content in an official context. Therefore it is important that the participants in this study match the profile of the intended user, namely a professional translator.

For these reasons, we recruited participants with an accreditation[10] in EN-GA translation and paid them an industry-consistent rate. We recruited four accredited translators, referred to from now on as P1, P2, P3 and P4 respectively.

Each participant was asked to post-edit 210 sentences: 10 practice sentences, 100 sentences translated using SMT and 100 sentences translated using NMT. The same source text was provided to all 4 translators. Figure 5.1 shows the distribution of MT output across participants. Two participants (P1 and P3) were presented with the SMT output using Job 1 data and the NMT output using Job 2 data (set-up A). The other two participants (P2 and P4) were asked to post-edit set-up B, consisting of Job 1 machine-translated using NMT and Job 2 machine-translated using SMT. Both set-up A and set-up B contain 10 practice sentences so that the translators can try out the PET environment without worrying about speed and familiarise themselves with the software. The output files from the practice segment are not

[8] `https://www.figure-eight.com/`

[9] `https://www.mturk.com/`

[10] The Foras na Gaeilge seal of accreditation for translators. Details of translators with this accreditation who are available on a part- or full-time basis are published on the Foras na Gaeilge website: `https://www.forasnagaeilge.ie/about/supporting-you/seala`

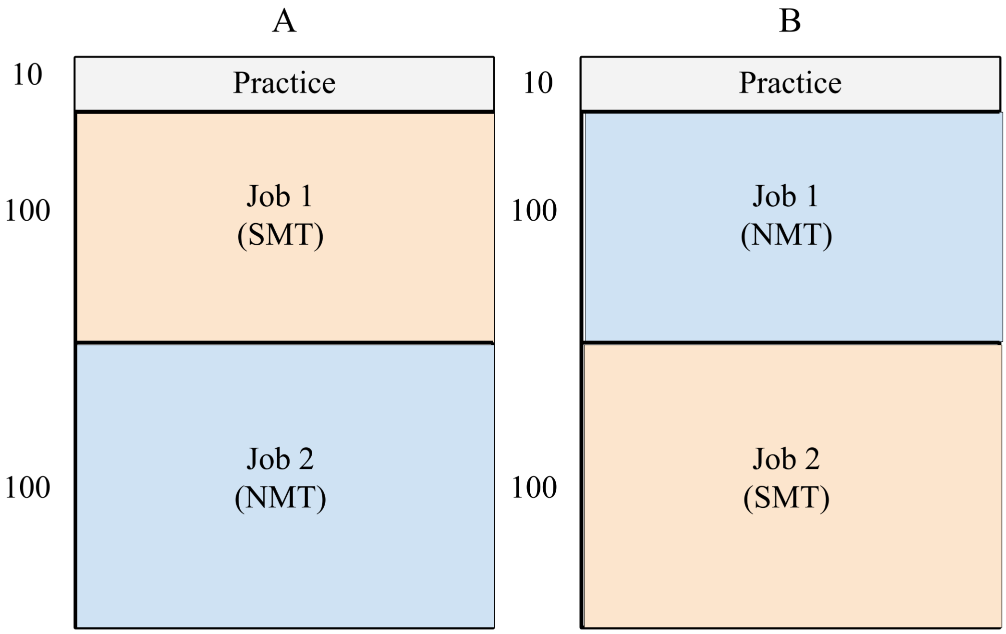


Figure 5.1: Distribution of MT output in PE tasks (not to scale)

be included in the results.

5.4.5 Survey questions

A post-task survey was implemented to gather information about the participants' experience and their opinions of the two MT outputs. Participants were not informed whether the MT output was produced by an SMT or NMT system. The survey was distributed via Google sheets. The following information was gathered in the survey:

- months/years experience as a professional translator (text box)
- months/years experience post-editing MT in a professional capacity (text box)
- view of MT as a tool to be used by professional translators (text box)
- which system seems most fluent (multiple choice)
  - System A
  - System B
  - No difference
- which system seems most accurate (multiple choice)

  - System A
  - System B
  - No difference

- which system the translator would prefer to post-edit (multiple choice)
  - System A
  - System B
  - No difference
- a text box for additional comments

5.5 Results and analysis

In this section we present the survey results and the results gathered via the PET output.

5.5.1 Survey results

The survey results show that all 4 participants are experienced translators. P1 has 25 years of professional translation experience, P2 5 years, P3 10 years part-time, and P4, 13 years. Two of the participants' (P2 and P4) have experience post-editing (PE) MT in a professional capacity, with 3 years (P2) and 5 years (P4) of PE experience each. This information is presented in Table 5.4.[11]

Participant	**Translator exp.**	**PE exp.**
P1	25 years	N/A
P2	5 years	3 years
P3	10 years†	N/A
P4	13 years	5 years

Table 5.4: Table displaying the amount of experience (exp.) each participant has a professional EN-GA translator and, if relevant, how much experience each has PE EN-GA output.

[11] In Table 5.4 a dagger (†) signifies that the experience is in a part-time capacity.

When asked for their views of MT as a tool to be used by professional translators, answers varied from positive ("It's a very useful tool") to cautious ("I think it depends very much on what the machine has been fed!"; "Improving constantly, but insufficient at present";) to negative ("It's not much use for English to Irish translation. It would take the same length of time to translate from scratch)." The positive but guarded responses came from participants with post-editing (PE) experience, whereas those without PE experience answered negatively. This may be an indication that there is a learning curve with PE before MT can be a valuable and useful addition to translation workflow.

Table 5.5 shows the survey results pertaining to differences between the two systems (SMT and NMT). The question "In general, which output did you perceive to be the most fluent-sounding?" is represented by the heading 'fluency'. 'Accuracy' is the heading used to represent the question "In general, which output did you did you perceive to be the most accurate in terms of semantics? (i.e. conveyed the meaning the best, fluency aside)." The final question dealing with SMT versus NMT, "Which output would you prefer to post-edit?" is represented with the heading 'prefer.' The participants were not aware which output was produced by which system; they were presented with two separate translation jobs.

Participant	**fluency**	**accuracy**	**prefer**
P1	NMT	NMT	No diff.
P2	No diff.	NMT	NMT
P3	No diff.	No diff.	No diff.
P4	NMT	NMT	NMT

Table 5.5: Survey responses relating to differences between SMT and NMT fluency, accuracy and participant preference.

It can be seen from Table 5.5 that 'NMT' and 'no difference' were the most common answers. Interestingly, none of the four participants gave 'SMT' as an answer to any of these questions. This contradicts results of previous work comparing EN-GA SMT and NMT using automatic metrics (see Chapter 3). It does, however, line up with the automatic metrics gathered during this study (BLEU, TER, ChrF and CharacTER scores suggested that the NMT output was of greater quality than that of SMT – see section 5.3 for more details).

5.5.2 Post-editing tool results

Once the participants had completed their PE tasks they sent the PET files back to us for analysis. The results gathered via the PET output provided us not only with the post-edited output, but also with the number of keystrokes, annotations, and seconds spent on each segment. We used this data to calculate the average seconds per segment, average keystrokes per segment, and the average unchanged segments per system per participant. These figures, as well as the human-targeted TER (HTER) scores (Snover et al., 2006), are displayed in Table 5.6.

participant	system	avg. time/seg.	avg. keys./seg.	avg unchanged segs.	HTER
1	SMT	102.4	91.47	0.12	0.33%
1	NMT	89.16	89.16	0.2	0.28%
2	SMT	119.86	207.09	0.11	0.52%
2	NMT	120.59	205.61	0.12	0.43%
3	SMT	173.15	90.44	0.17	0.36%
3	NMT	207.21	139.9	0.2	0.36%
4	SMT	193.06	100.49	0.1	0.43%
4	NMT	48.53	48.73	0.18	0.24%

Table 5.6: Table displaying the average (avg.) number of seconds (time) per segment (seg.), average number of keystrokes (keys.) per segments, average unchanged segments and HTER of each system for each participant.

Where MT for dissemination is concerned, temporal effort, or time spent post-editing, is arguably the most important metric as payment is usually based on words translated. Two of the four participants in this study (P1 and P4) were more productive when working with NMT output. This observation is inline with automatic metrics, as well as the survey responses from P1 and P4. The difference for P4 was sizeable (an average of 48.53 seconds per segment for NMT compared to 193.06 for SMT), although it should be noted that P4 was required to repeat the PE task for the NMT job due to a technical error. It is likely that this led to a faster PE time for this job, and that other values for this job are also skewed. P2 and P3 were more productive using SMT, although for P2 the difference is negligible (an average of 119.86 seconds per segment for SMT PE in comparison to 120.59 for

NMT). This contradicts survey results, where P2 expressed a preference for NMT and P3 reported no difference between the two systems.

HTER is a metric for evaluating MT output based on TER (described in Section 5.3). Using HTER, a human translator post-edits MT output and the score is calculated using the post-edit distance and the length of the reference. A low HTER score should equate to low PE effort, although in practice, post-editors may delete and retype text rather than taking the shortest possible route from raw MT to PE.

In the case of P1, P2 and P4, HTER was lower for NMT than SMT. Results from P3 showed negligible difference between the HTER of both systems (a difference of 0.0004). This matches the survey responses, with P1, P2 and P4 showing a preference to NMT and P3 reporting no difference between the SMT and NMT outputs.

In the survey, P1 reported that the NMT output was more fluent-sounding and more accurate. This is reflected in the data gathered from the PET output. From Table 5.6 we can see that P1 was quicker, used fewer keystrokes, and left more segments unchanged when post-editing NMT output. P1 did, however, choose ‘no difference’ when asked which output they would prefer to post-edit.

P2 also voiced a preference for NMT output over SMT output, although reported ‘no difference’ in fluency. Scores generated from PET data indicated little/no difference in time, keystrokes, and unchanged segments, although the HTER score was markedly improved for NMT.

Although P3 answered ‘no difference’ to all three questions comparing SMT and NMT, this is not reflected in the time and keystrokes, which indicated more favourable results for SMT, nor in the unchanged segments for which NMT had a higher score. It is, however, reflected in the HTER scores which are almost identical for both outputs.

P4 reported NMT to be more fluent sounding, more accurate, and the output they would most prefer to post-edit. This is reflected in all metrics present in Table 5.6, where the results for the NMT output show a marked improvement over those of the SMT output, apart from the number of unchanged segments. However, as mentioned in Section 5.5.1, P4 had to repeat the entire PE task for the NMT output. This may have lead to a faster PE time with fewer keystrokes and, relatedly, a lower HTER score.

Overall, these results suggest HTER to be the metric that aligns best with translators' expertise, rather than the other scores we gathered automatically, in that when a translator preferred NMT over SMT, or had no preference, the HTER reflected this.

5.5.3 Analysing post-edited output

With both the survey responses and figures generated using results from PET varying substantially from translator to translator, we chose to take a closer look at the differences in PE output provided by the four participants. To identify potentially interesting sentences, we used compare-mt (Neubig et al., 2019), a tool designed to analyse MT output and provide the user with sentences which differ greatly. Although human-generated translations are not the intended input for compare-mt, it was still useful in identifying cases where the participants gave different translations.

Input:	If you have been allocated as a decision-maker..
SMT:	Má tá tú mar a déantóir cinntí..* *If you are a decision manufacturer..*
P1:	Más cinnteoir thú air.. *If you are a decision-maker for it..*
P3:	Má ainmníodh thú mar chinnteoir.. *If you are named as a decision-maker..*
NMT:	Má roghnaíodh mar chinnteoir thú.. *If you are chosen as a decision-maker..*
P2:	Má shanntar ról mar chinnteoir ort.. *If the role of decision-maker is assigned to you..*
P4:	Má roghnaíodh mar chinnteoir thú.. *If you are chosen as a decision-maker..*

Table 5.7: A portion of the PE output from P1, P2, P3 and P4.

Table 5.7 shows a shortened portion of a segment of post-edited output produced by P1, P2, P3 and P4.[12] It can be seen, even to those who do not speak Irish, that all four translators chose to post-edit the MT input in a different way. In fact, there is no word that is repeated (with the same inflections) throughout all four translations. Despite all four translations being correct, it stands to reason that automatic values generated for this output, such as HTER and number of keystrokes, would also differ considerably. This highlights the limitations of such metrics, as well as the need for multiple references when generating accurate and meaningful automatic evaluation scores.

Source	to ensure.. in the **FOI** legislation..
SMT:	chun a chinntiú.. sa reachtaíocht um **Shaoráil Faisnéise**.. *to ensure.. in legislation surrounding the **Freedom of Information..***
P1:	cinntiú.. sa reachtaíocht um **Shaoráil Faisnéise**. *ensure.. in the legislation surrounding the **Freedom of Information..***
P3:	chun a chinntiú.. sa reachtaíocht **SF**. *to ensure.. in the **FOI** legislation..*
NMT:	féachaint.. sa reachtaíocht um **Shaoráil Faisnéise**.. *see.. in the legislation surrounding the **Freedom of Information..***
P2:	a fheacháint.. i reachtaíocht um **Shaoráil Faisnéise..** *to see.. in legislation surrounding the **Freedom of Information..***
P4:	féachaint.. sa reachtaíocht um **Shaoráil Faisnéise..** *see.. in the legislation surrounding the **Freedom of Information..***

Table 5.8: A portion of the PE output from P1, P2, P3 and P4.

Similarly, in Table 5.8, all four participants chose slightly different translations of the source text.[13] In this example, the importance of context can be seen. In the source text, the acronym for Freedom of Information (FOI) is not expanded. Despite this, only P3 chooses to use the equivalent Irish acronym, possibly due to both MT systems producing the expanded acronym (shown in bold). The three other translators (P1, P2 and P4) chose to preserve the expanded acronym in the GA PE sentence. It could be the case that, in Irish, the acronym is not as instantly recognised as its English counterpart. This is quite common, when an acronym

[12] In Table 5.7 the EN data provided to the translators as the source text is also provided. The relevant MT output provided to translators is given above the participants output. A gloss for each sentence is indicated in italics below each GA output. An asterisk (*) indicates that the segment is not grammatically correct.

[13] In Table 5.8, the EN data provided to the translators as the source text is also provided. The relevant MT output provided to translators is given above the participants output. A gloss for each sentence is indicated in italics below each GA output.

is commonly used in one language but not in another. Without training data to reflect this, it is unlikely that an MT system would produce such an output. This inconsistent spelling-out of the acronym in the post-edited texts again indicates the importance of in-domain training data and of seeking the advice of professional translators when selecting data to train MT systems.[14]

5.6 Summary and Conclusions

In this Chapter, we have presented the first human evaluation study for EN-GA SMT and NMT. We have shown that, while automatic metrics can be useful in obtaining a rough idea of MT system quality, they do not always correlate with human evaluation. Although, according to automatic metrics, NMT was identified as the 'better' system and was the system translators deemed most accurate – three of the four translators chose the NMT system as the most accurate output in the post-task survey (see Table 5.5) – this did not consistently align with the scores generated using the PET output or with the translators' perceptions of fluency or the system which they would most prefer to post-edit (two of the four translators chose 'no difference' for both 'fluency' and 'prefer' in Table 5.5.).

Overall, we can see that, even with just four participants, results can vary from translator to translator in terms of survey results as well as results gathered as a direct outcome of PE. Therefore, if experienced professional human translators do not produce the same translations or PE data, it is unreasonable to expect any one automatic metric to perfectly mirror human evaluation.

In this study, we have observed HTER as the metric which most closely matches our participants' survey responses. However, it is important to note that with this study being limited to four participants we are unable to make definitive conclusions as to the best metric with which to guide EN→GA MT system development. As might be expected, the recommended approach would be to use human evaluation

[14] It is also probable that in a professional translation environment, especially in public administration, style guidelines may address whether an acronym should be expanded or not.

wherever possible, and, in cases where this is not feasible, a combination of automatic metrics will provide the broadest snapshot of MT quality.

In terms of future work, we propose a similar study with more participants. We have seen that translators vary in MT PE approaches, experience and opinion. Accordingly, more participants would provide us with a more accurate picture of EN→GA MT quality and would provide us with a greater number of data points to extrapolate from. We also suggest a more fine-grained evaluation of EN→GA MT output. In this study, we elicit opinions of MT quality over 100-sentence documents in general. In the future it may be beneficial to examine specific differences between EN-GA SMT and NMT at the sentence-level, examining variations in errors in case, semantics, tense, etc.

In terms of type of MT system on which to focus EN-GA MT development efforts, although we have highlighted that not all results were consistent, for the most part the automatic metrics and, more importantly, the human translators, veered toward NMT over SMT. Both automatic metrics and (for the most part) human evaluation results have shown NMT to be the better system with these particular parameters and data, although the difference is minimal in places. We predict that, with the derogation of the production of Irish language text within the EU ending in 2021, the greater amount of training data will more favourably affect NMT than SMT and the differences between the two will become more pronounced. Experiments in Chapter 3 indicated that NMT automatic scores were continuing to increase with the addition of data, whereas the SMT scores were beginning to plateau. The ending of the derogation also means that we have a greater need than ever for EN-GA MT systems designed with the end-user in mind. Therefore, we recommend furthering the investigation of EN-GA NMT as this is the area more likely to produce results that suit the needs of human translators.

Chapter 6

Conclusion

In this thesis, we have addressed gaps in research of EN→GA MT by exploring (i) the limited data publicly available for EN→GA MT experimentation, and (ii) the challenges that exist when applying state-of-the-art statistical and neural MT approaches to this language pair. We have gathered and curated sufficient datasets to facilitate the training of good quality EN-GA SMT and NMT systems, and have experimented with a hybrid approach to combine the two paradigms. We have tested the output of these systems, not only with automatic metrics but also with the first published human evaluation study of EN-GA SMT and NMT.

In this chapter, we summarise the work we undertook to address our research questions (Section 6.1) and list our contributions to the field (Section 6.1.1). We provide some possible avenues of future work (Section 6.2), based on this research, and end with some final remarks (Section 6.3).

6.1 Research questions

In this section, we revisit the research questions introduced in Chapter 1 and discuss the extent to which the research presented in this thesis has answered each question.

- **RQ1:** *What are the existing corpora available for use in EN-GA MT? Given related projects in this area, how can we increase the size of this corpus?*

- **RQ2**
  - **(a):** *How can improvements in the quality of EN-GA statistical machine translation (SMT) and neural machine translation (NMT) be captured in terms of automatic evaluation metrics?*
  - **(b):** *How do these SMT and NMT systems compare; do the different systems produce the same types of errors? If so, would a hybrid SMT-NMT system outperform both baselines? Would this be confirmed in experiments on a very closely-related language pair?*
  - **(c):** *How can improvements in the quality of EN-GA statistical machine translation (SMT) and neural machine translation (NMT) be captured in terms of human evaluation? To what extent do the findings from the human evaluation corroborate the findings from the automatic evaluations?*

RQ1 Our first goal was to identify the existing corpora available for use in Irish MT and to explore if additional corpora could be gathered. Chapter 2 described the publicly available data, as well as methods for supplementing this dataset via web-crawling, direct contact with various national organisations and, to a lesser extent, crowd-sourcing. We assembled from public data 40,364,626 words of GA data, and added an additional 23,486,968 words through additional data collection efforts.

RQ2 Our second goal was to investigate how to develop good quality EN-GA SMT and NMT systems, and to what extent human and automatic metrics captured this quality. The datasets outlined in Chapter 2 were used to train SMT and NMT systems, with the results presented in Chapter 3. Results indicate that with a domain-specific test set, unsurprisingly, the domain of the training data is an important factor and has an impact on automatic evaluation metrics.

In Chapter 3 we compared the quality of SMT and NMT systems, with evaluation via automatic scores. We show that system tailoring can improve on baseline scores with both SMT and NMT. When scores for each system were compared, we saw that a tailored SMT system (BLEU: 39.7) outscores an out-of-the-box NMT system (BLEU: 33.03), but with the addition of more data and implementation of a Transformer architecture, NMT scores (BLEU: 42.66) outperform that of corresponding SMT systems (BLEU: 40.55).

Regarding the sub-question '*would a hybrid SMT-NMT system outperform both baselines?*', we explored backtranslation as a method of building a hybrid MT system. Ultimately, while the results did not show positive improvements in automatic scores we contend that a repeated experiment with a different set-up (e.g. using the Transformer architecture as the basis for the NMT system and choosing monolingual data that is closer in domain to the domain of the test set) could show more beneficial results. Furthermore, as linguistic analyses gained through the generation of sentence-level BLEU scores showed, automatic metrics do not necessarily represent MT output quality.

In terms of comparing EN-GA SMT and NMT via human translation, we described a manual evaluation study in Chapter 5. We elicited the expertise of professional translators to post-edit SMT and NMT output and provide insights as to the accuracy, fluency of each system, and which system they would prefer to post-edit. A combination of survey results and metadata acquired through professional translator post-editing (e.g. temporal effort, number of keystrokes) revealed that NMT appears to be at least as good as SMT, with some scores showing NMT to have produced better quality output. Although we would recommend a larger, more detailed study in the future, as the first published human evaluation study of EN-GA MT, it gives previously unseen insights into the quality of EN-GA SMT and NMT.

6.1.1 Contributions

Using the research questions to shape the research presented in this thesis, we have provided, to some extent, a guide for the training of EN-GA SMT and NMT, including the gathering/curation of appropriate data and analysis of output via human and automatic metrics. The main contributions can be summarised as follows:

- We provided and presented the first published paper on EN-GA SMT and resources (Dowling et al., 2015). Results were promising, given a lack of training data available at that time.

- We carried out the first known comparison of EN-GA SMT and NMT through automatic metrics and a preliminary linguistic analysis (Dowling et al., 2018). Results showed that, although SMT scored higher in terms of automatic metrics, there were still examples of NMT outperforming SMT. This motivated us to continue to explore ways of improving EN-GA NMT.

- With the goal of EN-GA MT to be used as a tool for professional translators, it is important that MT development be user-driven. To facilitate this, we carried out the first published human evaluation study of EN-GA MT. Insights gained via translator surveys, combined with those gleaned from recorded background data indicated that NMT could be a more promising avenue of research than SMT for the EN-GA language pair. This study will be important for advising developers of GA MT at a national and European level, where translator productivity and satisfaction is important. This is underlined by the significant societal impact achieved by the use of the MT system produced in the Tapadóir project as a productivity tool for translators in the DCHG and has a potential use in other public bodies.

- We applied backtranslation, not just to EN-GA MT, but also to the Scottish Gaelic MT, a language which has even less publicly-available data than Irish. We show through this study that our work can be useful in informing low-resource MT research for other languages, in particular those in the Celtic language family.

- We have participated in EU-related data collection initiatives and played a part in the collection of monolingual and bilingual corpora for use in the development of EN-GA MT systems at a national and European level.

6.2 Future work

In terms of future work, there are many ways in which this research could be continued.

In Chapter 2, data collection efforts are outlined. This is not, however, an exhaustive list of all Irish data that will ever exist. Irish content and translations will continue to be published both at a national and European level. As such, data collection efforts can and should continue. It is our hope that language data management practices will continue to improve, in order to facilitate continued data collection for GA MT (as per Lynn et al. (2019)). At present, the EU-funded PRINCIPLE project is gathering Irish data in the domain of e-Justice and e-Procurement in order to improve the quality of the DGT's eTranslation system. With the derogation on the production of Irish content coming to an end in 2021, the amount of Irish data available publicly will increase hugely. This is expected to have a positive effect on EN-GA MT, as per the virtuous cycle of data collection and MT outlined in Chapter 1. In terms of future data collection via web-crawling, we also recommend experimenting with the Bitextor crawler (Esplá-Gomis and Forcada, 2009).

Chapter 3 described various methods for improving EN-GA SMT and NMT. MT research is a quickly-evolving field, with new state-of-the-art methods being published constantly. One avenue of research is to enhance MT with dependency-aware self-attention, as in Bugliarello and Okazaki (2020). This method uses Transformer, which we have shown to have promising results for EN-GA MT, and has been tested on the less-resourced English-Turkish language pair, which leads us to view it as a promising method for improving EN-GA MT.

With EN-GA SMT seemingly outperforming NMT, we advise research efforts to focus on NMT. With Irish as a less-resourced language, all data gathered is valuable. To fully utilise the data available, we recommend further investigation into the addition of monolingual data in EN-GA NMT. Zhou and Keung (2020) present a method of leveraging monolingual data to improve NMT and test using the English-Romanian language pair. Siddhant et al. (2020) recommend using monolingual data, in combination with self-supervision to improve the quality of low-resource language MT, within the context of multilingual MT. This could have a particular use-case of MT developers aiming to provide MT for official EU languages, including Irish. Monolingual data has also shown to be useful in the training of MT systems with extremely low amounts of parallel text, via unsupervised MT (e.g. Artetxe et al. (2019)). An unsupervised EN-GA MT system, perhaps in the context of hybrid MT, could be a promising direction of future research.)

We explored backtranslation as a method for combining SMT and NMT systems in Chapter 4. Although this did not show a positive increase in terms of automatic evaluation scores, we make recommendations for repeating these experiments. Firstly, we recommend experimenting with different sources of monolingual data as a basis for the artificial parallel dataset. Furthermore, with NMT systems trained using Transformer showing the most promise, we recommend training the NMT portion of the hybrid system using Transformer instead of the generic OpenNMT system described in Chapter 4.

Moreover, the backtranslation engine is kept static in the experiments we performed. In the future, we recommend retraining it with incrementally more amounts of authentic data at each iteration. Finally, with the expected increase in NMT quality, aligned with the increase in data, it might be beneficial to consider two different NMT systems as the basis for backtranslation, rather than one SMT and one NMT system.

Other experiments described in Chapter 4 show backtranslation to be useful in the training of GD↔GA and GD↔EN MT systems. We would be interested in repeating these experiments with other low-resource Celtic languages such as Welsh or Manx.

Chapter 5 provides the first published human evaluation study of EN-GA SMT and NMT. While it is a valuable contribution, there is much scope for expansion. Our study provided insights from just 4 professional translators due to time and budget constraints. An expanded study with more translators and a more fine-grained analysis of MT output would provide a more detailed insight into actual perceived MT quality and would be better equipped to inform MT developers which particular areas need most attention.

We would also recommend follow-on experiments with the metrics gathered, such as studying the the correlation of sentence-level metrics such as TER with post-editing time, computing more statistics about post-edited segments (e.g. average sentence length or total number of words) and computing time per character and time per word in addition to the time per segment reported. Such statistics could then be used to compare productivity in relation to a translators' daily throughput (generally taken as between 2500 and 5000 words a day).

6.3 Final remarks

Finally, we highlight that MT development, in particular development of Irish MT, should be user-driven. The main motivation for this research has been to aid national and European bodies to meet the translation needs of Irish language speakers. We have done this through the gathering and curation of suitable language data resources, in co-operation with EU-led projects as well as the investigation of methods of improving EN-GA SMT and NMT, a previously neglected language pair within MT research.

With the derogation on the production of Irish-language documents within the EU due to lift in 2021, there will be increased pressure on an already stretched pool of professional translators. It is important for MT developers to consider the needs and recommendations of these professional translators, rather than relying solely on automatic metrics to guide development. Ethical considerations surrounding MT use should therefore be at the forefront of development.

Although there is huge room for improvement regarding the current quality of EN-GA MT, we hope that the research presented in this thesis can be a useful springboard for the continued development of user-aligned, fit-for-purpose MT systems.

Appendix A

List of Web-crawled websites

Name of organisation	URL
Logainm (Fiontar, DCU)	http://www.logainm.ie/
Abair (TCD)	http://www.abair.tcd.ie/
An Puball Gaeilge	http://www.anpuballgaeilge.com/
Dept. of Education and Skills	http://education.ie/
Oideas Gael	http://www.oideas-gael.com/
Seachtain na Gaeilge (Conradh na Gaeilge)	http://snag.ie/
Údarás na Gaeltachta	http://www.udaras.ie/en/
Met Éireann	http://met.ie/
Fás	http://www.fas.ie/
Irish Government	http://www.gov.ie/
Galway County Council	http://www.galway.ie/en/
Dept. of Housing, Planning and Local Government	http://www.housing.gov.ie
Central Applications Office	http://www.cao.ie/
Companies Registration Office	https://www.cro.ie/
eFlow	https://www.eflow.ie/
Gaelscoil Chaladh an Treoigh	http://www.gaelcat.com/
Gaelchultúr	http://gaelchultur.com/
Gaelscoil Nás	http://gaelscoilnas.com/
Gael-Taca	http://www.gael-taca.com
Gaoluinn	http://www.gaoluinn.com/
Glór na nGael	http://www.glornangael.ie/
Líofa	https://www.liofa.eu/
Acts of the Oireachtas	http://achtanna.ie/
Ainm	http://www.ainm.ie/
Motor Tax Online	https://www.motortax.ie/
National Museum of Ireland	http://www.museum.ie/
Marine Institute	http://www.marine.ie/
Galway le Gaeilge	http://www.gleg.ie
Scríobh (resource portal)	http://www.scriobh.ie/
Clare County Council	http://www.clarecoco.ie

Appendix B

Crowd-sourcing: instructions for participants

B.1 Translation guidelines

The following are the translation guidelines provided to users to aid them in their translation.

- **Placeholders:** #hashtags and @twitterhandles are to be left untranslated. Emoticons have been replaced by the placeholder [emoticon]. Please retain these placeholders in your Irish translation (or English translations) also.
  e.g. My Dad [emoticon] soaked but smiling #ge16 → *M'athair [emoticon] fliuch báite ach fós gealgháireach #ge16*

- **Case:** Please keep translations case sensitive where possible.
  e.g.: FULL HOUSE Great night tonight @SorchaNicC #GE16 launch. → *TEACH LÁN Oíche iontach anocht ag seoladh #GE16 @SorchaNicC.*

- **Text speak:** Where possible, please translate English text speak to Irish text speak (and vice versa), where there are equivalents.
  e.g. tnx (thanks) → *grma (go raibh maith agat).* If there is no shortened Irish/English equivalent that you are aware of, translate the word into its full form.

- **Tweet length:** Although the English tweets have been limited to 140 characters, your Irish translations do not have to adhere to this.

- **Pre-translate options:** It is acceptable to use Google Translate to pre-translate the tweets and correct the output – if you find it helpful. If it is too much of a hindrance, translation from scratch might work better. Note that the translations do not have to be 100% sound. Remember that the quality of Twitter language is questionable at the best of times, so your best shot is enough. Where there is ambiguity, go with your intuitive translation.

- **Confidence level:** After having translated the tweet, you are asked to indicate how confident you are that your translation is accurate. Please rate your translation on a scale of 1–10 from the drop-down menu provided.

- **Skip translation:** If you want to skip a tweet leave the translation field blank and submit a confidence level of 0

Appendix C

Human evaluation study – guidelines for participants (GA)

C.1 Ag cur tús leis

1. Bí cinnte go bhfuil java ar do ríomhaire. Mura bhfuil, is féidir é a íoslódáil anseo: `https://java.com/en/download/`

2. Íoslódáil an ceangaltán 'participant.zip' ón gcomhad Drive

3. Dízipeáil an ceangaltán

4. Má oibríonn do chóras oibriúcháin ar Windows, cliceáil faoi dhó ar an gcomhad darb ainm 'run.bat'

5. Má tá tú ar Linux, scríobh ./run.sh i líne na n-orduithe (command line)

6. Seo mar a bhreathnóidh an clár (nuair a osclófar é – Fig. C.1):

7. Brúigh ar 'practice.pej'

8. Brúigh ar 'Start'

9. Beidh an scáileán mar seo (Fig. C.2):

10. Brúigh ar an mbosca le Gaeilge istigh ann chun é a chur in eagar

11. Brúigh ar $\Downarrow$ chun an chéad abairt eile a fháil

12. Brúigh ar $\underline{\Downarrow}$ chun an chéad abairt eile nach bhfuil aon athrú déanta air a fháil

13. Brúigh ar $\Uparrow$ chun an abairt roimhe a fháil

14. Brúigh ar $\overline{\Uparrow}$ chun an abairt is gaire thuas nach bhfuil aon athrú déanta air a fháil

15. Ní gá duit na habairtí a aistriú in ord

16. Is féidir leat dul ar ais go dtí aistriúchán agus é a athrú arís

17. Nuair a bheidh tú críochnaithe brúigh ar an deilbhín sábhála

18. Brúigh ar deilbhín an dorais bhig chun críochnú

19. Nuair atá tú críochnaithe le 'practice.pej', lean ar aghaidh le 'job1.pej' and déan an rud céanna

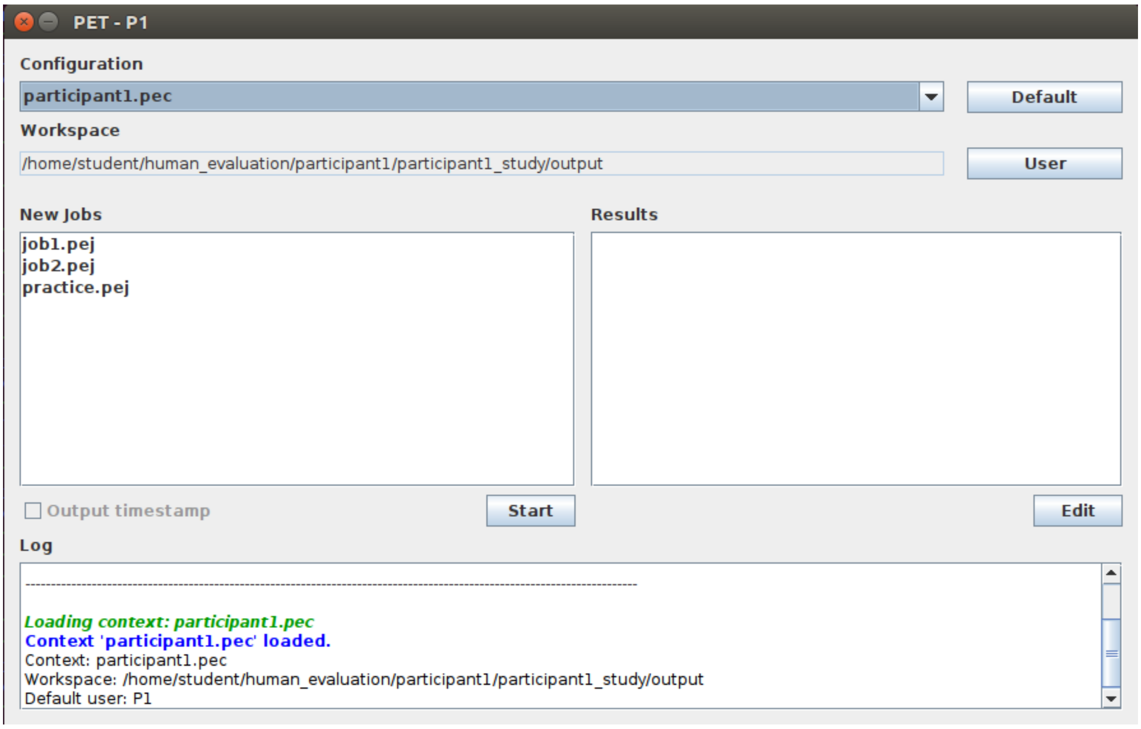
20. Nuair atá tú críochnaithe le 'job1.pej', lean ar aghaidh le 'job2.pej'

21. Ní chaithfidh tú an rud ar fad, nó an jab ar fad, a dhéanamh in aon iarracht amháin

22. Tá níos mó treoracha ar fáil ar líne, anseo: `http://wilkeraziz.github.io/dcs-site/pet/manual/r134.pdf`

C.2 Ag leanúint ar aghaidh ó tréimhse oibre luaithe

- Brúigh ar an gcomhad oiriúnach faoi 'Results' agus ansin ar 'Edit'
- Beidh dath glas ar na cinn a chuir tú in eagar roimh ré
- Beidh dath dearg ar na cinn nár chuir tú in eagar go fóill
- Ná brúigh ar an gcomhad faoi 'new jobs'! Beidh an chuid a rinne tú níos luaithe imithe!



Configuration

participant1.pec

Workspace

/home/student/human_evaluation/participant1/participant1_study/output

New Jobs

job1.pej
job2.pej
practice.pej

Results

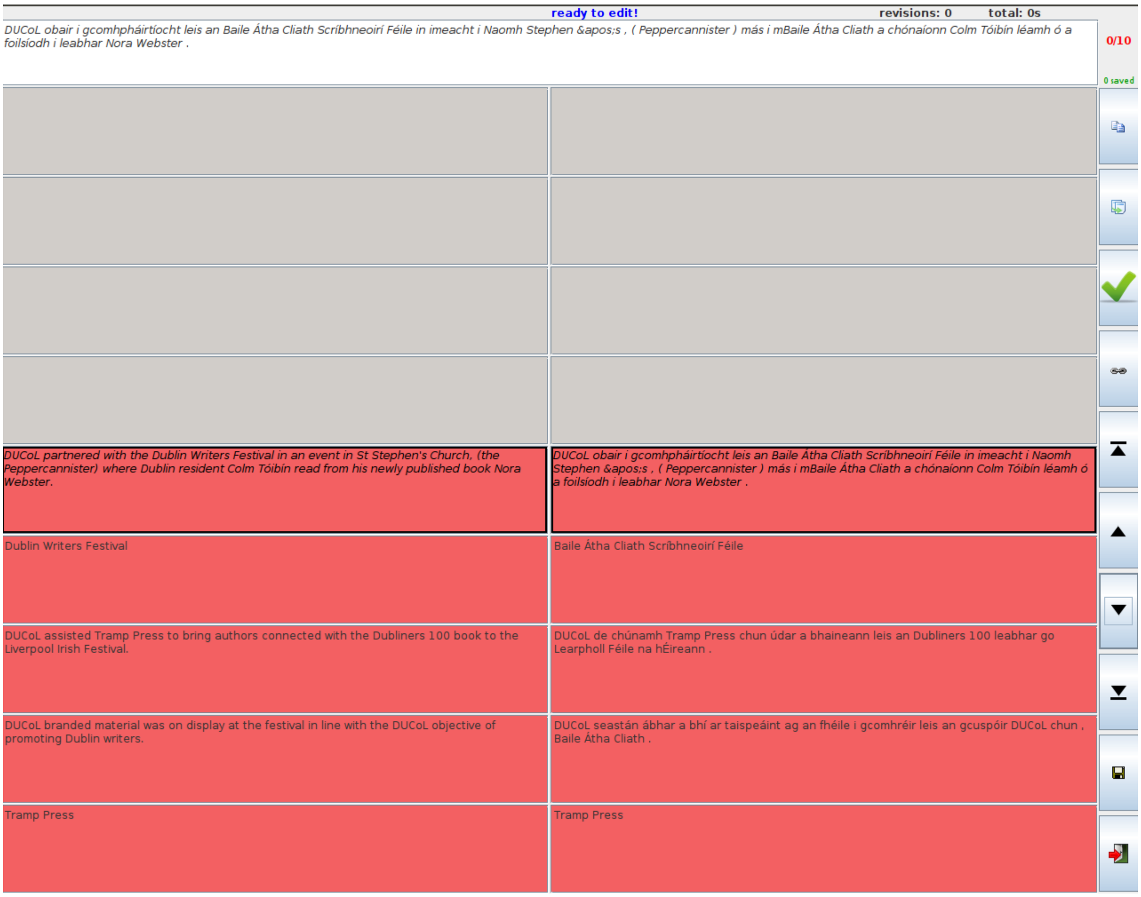
□ Output timestamp

Log

```
----------------------------------------------------------------------------------------------------------------

Loading context: participant1.pec
Context 'participant1.pec' loaded.
Context: participant1.pec
Workspace: /home/student/human_evaluation/participant1/participant1_study/output
Default user: P1
```

Figure C.1: An uirlis PET nuair a osclófar é



ready to edit! revisions: 0 total: 0s

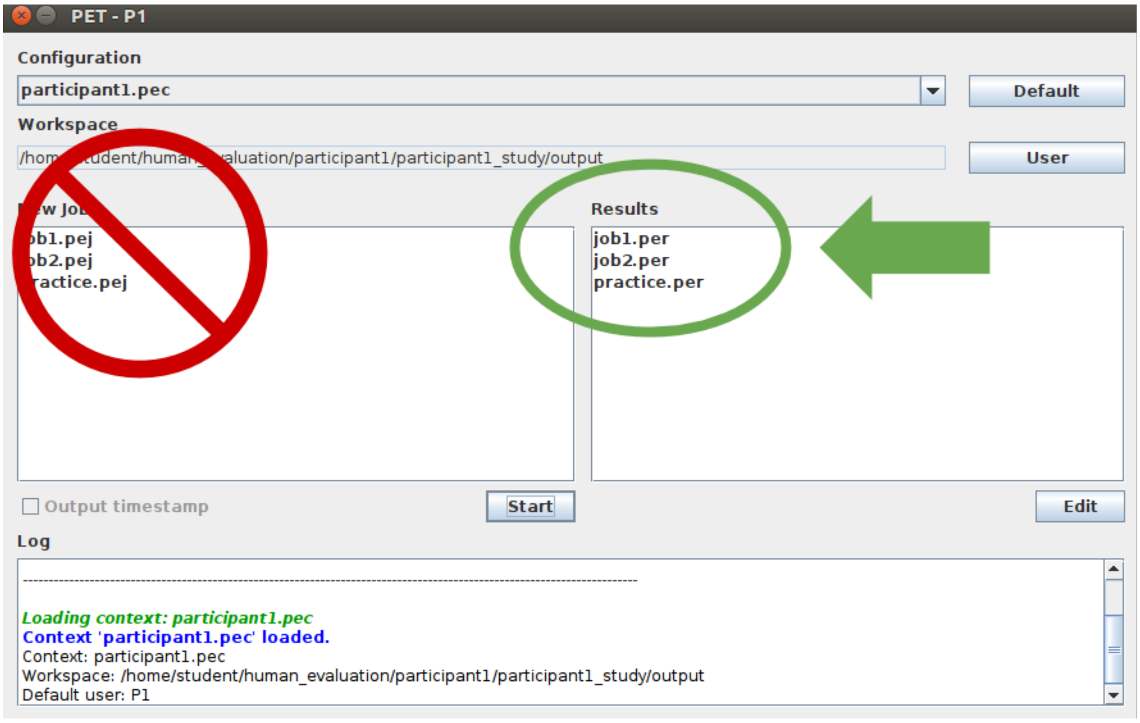
DUCoL obair i gcomhpháirtíocht leis an Baile Átha Cliath Scríbhneoirí Féile in imeacht i Naomh Stephen &apos;s , (Peppercannister) más i mBaile Átha Cliath a chónaíonn Colm Tóibín léamh ó a foilsíodh i leabhar Nora Webster .

0/10

0 saved

DUCoL partnered with the Dublin Writers Festival in an event in St Stephen's Church, (the Peppercannister) where Dublin resident Colm Tóibín read from his newly published book Nora Webster.	*DUCoL obair i gcomhpháirtíocht leis an Baile Átha Cliath Scríbhneoirí Féile in imeacht i Naomh Stephen &apos;s , (Peppercannister) más i mBaile Átha Cliath a chónaíonn Colm Tóibín léamh ó a foilsíodh i leabhar Nora Webster .*
Dublin Writers Festival	Baile Átha Cliath Scríbhneoirí Féile
DUCoL assisted Tramp Press to bring authors connected with the Dubliners 100 book to the Liverpool Irish Festival.	DUCoL de chúnamh Tramp Press chun údar a bhaineann leis an Dubliners 100 leabhar go Learpholl Féile na hÉireann .
DUCoL branded material was on display at the festival in line with the DUCoL objective of promoting Dublin writers.	DUCoL seastán ábhar a bhí ar taispeáint ag an fhéile i gcomhréir leis an gcuspóir DUCoL chun , Baile Átha Cliath .
Tramp Press	Tramp Press

Figure C.2: An uirlis PET le linn eagarthóireachta



Configuration: participant1.pec

Workspace: /hom tudent/huma aluation/participant1/participant1_study/output

ew Jo
bb1.pej
bb2.pej
ractice.pej

Results
job1.per
job2.per
practice.per

Log

```
--------------------------------------------------------------------------------------------------------------------

Loading context: participant1.pec
Context 'participant1.pec' loaded.
Context: participant1.pec
Workspace: /home/student/human_evaluation/participant1/participant1_study/output
Default user: P1
```

Figure C.3: Conas leanúint ar aghaidh ó tréimhse oibre luaithe

C.3 Stíl

- Ceartaigh an cás (ceannlitreacha nó cás beag) má tá sé mícheart
- Is féidir an t-aschur MT a scriosadh agus an abairt a aistriú ón dtús mura féidir leat an t-aschur a chur in eagar
- Athraigh an téacs go dtí go bhfuil ardchaighdeán Gaeilge air - an caighdeán céanna a bheadh air dá mbeifeá ag aistriú ó bhonn
- Bí cinnte go bhfuil an bhrí chéanna idir an Ghaeilge agus an Béarla

C.4 Rialacha an staidéir

- Is féidir foclóir (fisiciúil nó ar líne) a úsáid
- Ní féidir aon uirlis aistriúcháin eile a úsáid, mar shampla Google Translate nó Bing Translate
- Ní féidir aon chuimhne aistriúcháin (translation memory) a úsáid

- Ní féidir leat bheith i dteagmháil leis na rannpháirtithe eile faoin staidéar seo agus é fós ar siúl

C.5 Ginearálta

- Is staidéar anaithnid é seo ach más mian leat d'ainm a bheith liostaithe i nótaí buíochais an pháipéir, seol r-phost chuig Meghan ag meghan.dowling@adaptcentre.ie
- Seol r-phost chuig Meghan má tá ceist nó fadhb agat: meghan.dowling@adaptcentre.ie

Appendix D

Human evaluation study –guidelines for participants (EN)

D.1 Beginning

1. Make sure java is installed on your computer. If not, it can be downloaded here: `https://java.com/en/download/`

2. Download the attachment MT.zip from the Drive folder

3. Unzip the attachment

4. If your operating system works on Windows, double click on the file called run.bat

5. If you're on Linux, write ./run.sh in the command line

6. The program (when opened) will look like so: (see Fig. D.1)

7. Click 'practice.pej'

8. Click 'Start'

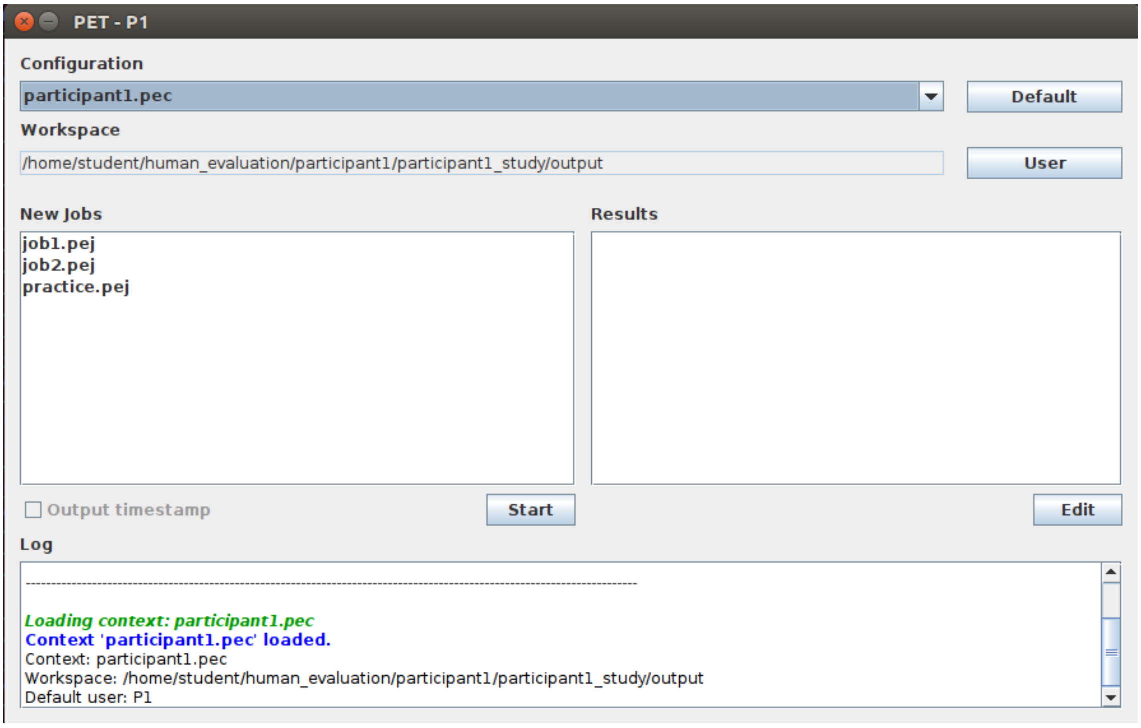
9. The screen will look like this: (see Fig. D.2)

10. Press on the box with Irish inside to edit it

11. Press the down arrow to get the next sentence
12. Press on the down arrow with a line to find the next unchanged sentence
13. Press on the up arrow to get the previous sentence
14. Press on up arrow with a line to get to the nearest sentence above that has not been changed
15. You do not have to translate the sentences in order
16. You can go back to a previous translation and change it again
17. When finished press the save icon
18. Press on the icon of a small door to exit
19. When you are finished with ‘practice.pej’, continue to ‘job1.pej’ and follow the same steps.
20. When you are finished with ‘job1.pej’, continue to ‘job2.pej’
21. You don’t have to complete the whole thing, or an entire job, in one session
22. More guides are available online, here: `http://wilkeraziz.github.io/dcs-site/pet/manual/r134.pdf`

D.2 Continuing from an earlier work session

- Click on the appropriate file under ‘Results’ and then on ‘Edit’
- Segments you have edited previously will be in green
- Those you haven’t edited will be red
- Don’t click on the file beneath ‘new jobs’! Everything you’ve done until now will be gone!



Configuration: participant1.pec

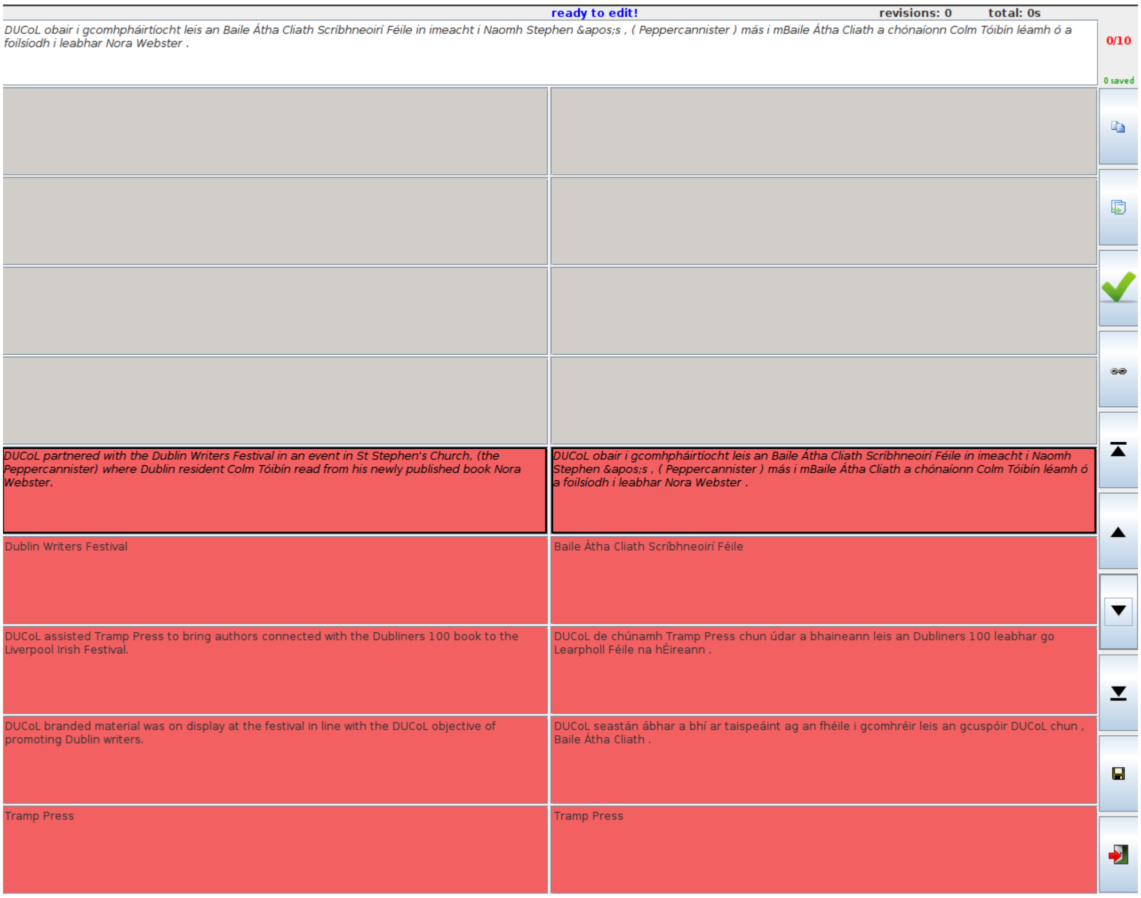
Workspace: /home/student/human_evaluation/participant1/participant1_study/output

New Jobs:
job1.pej
job2.pej
practice.pej

```
------------------------------------------------------------------------------------------------------------------

Loading context: participant1.pec
Context 'participant1.pec' loaded.
Context: participant1.pec
Workspace: /home/student/human_evaluation/participant1/participant1_study/output
Default user: P1
```

Figure D.1: The PET environment when it is opened



DUCoL obair i gcomhpháirtíocht leis an Baile Átha Cliath Scríbhneoirí Féile in imeacht i Naomh Stephen &apos;s , (Peppercannister) más i mBaile Átha Cliath a chónaíonn Colm Tóibín léamh ó a foilsíodh i leabhar Nora Webster .

DUCoL partnered with the Dublin Writers Festival in an event in St Stephen's Church, (the Peppercannister) where Dublin resident Colm Tóibín read from his newly published book Nora Webster.	*DUCoL obair i gcomhpháirtíocht leis an Baile Átha Cliath Scríbhneoirí Féile in imeacht i Naomh Stephen &apos;s , (Peppercannister) más i mBaile Átha Cliath a chónaíonn Colm Tóibín léamh ó a foilsíodh i leabhar Nora Webster .*
Dublin Writers Festival	Baile Átha Cliath Scríbhneoirí Féile
DUCoL assisted Tramp Press to bring authors connected with the Dubliners 100 book to the Liverpool Irish Festival.	DUCoL de chúnamh Tramp Press chun údar a bhaineann leis an Dubliners 100 leabhar go Learpholl Féile na hÉireann .
DUCoL branded material was on display at the festival in line with the DUCoL objective of promoting Dublin writers.	DUCoL seastán ábhar a bhí ar taispeáint ag an fhéile i gcomhréir leis an gcuspóir DUCoL chun , Baile Átha Cliath .
Tramp Press	Tramp Press

Figure D.2: The PET environment during editing

D.3 Style

- Correct the case (initials or small case) if it is wrong
- The output MT can be deleted and the sentence translated from scratch if you cannot edit the output
- Change the text until it has a high standard of Irish - it should be the same standard as if you were translating from scratch
- Make sure that the Irish and English have the same meaning

D.4 Rules of the Study

- A dictionary (physical or online) can be used
- No other translation tool can be used, such as Google Translate or Bing Translate
- No translation memory can be used
- You may not be in contact with the other participants about this study while it is ongoing

D.5 General

- This is an anonymous study but if you would like to be listed in the paper's acknowledgements please email Meghan at meghan.dowling@adaptcentre.ie
- Please email Meghan if you have a question or problem: meghan.dowling@adaptcentre.ie

Bibliography

Afli, Haithem, Sorcha McGuire, and Andy Way (2017). “Sentiment translation for low resourced languages: Experiments on irish general election tweets”. In: *Proceedings of the 8th International Conference on Intelligent Text Processing and Computational Linguistics*. Budapest, Hungary.

Ambati, Vamshi and Stephan Vogel (2010). “Can crowds build parallel corpora for machine translation systems?” In: *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon’s mechanical turk*. Los Angeles, U.S.A., pp. 62–65.

An Coimisinéir Teanga (2003). *Official Languages Act 2003 Guidebook.*

— (2020). *An Coimisinéir Teanga homepage.* URL: `https://www.coimisineir.ie/Failte`.

Aranberri, Nora, Gorka Labaka, Arantza Díaz de Ilarraza, and Kepa Sarasola (2017). “Ebaluatoia: crowd evaluation for English–Basque machine translation”. In: *Language Resources and Evaluation* 51.4, pp. 1053–1084.

Arcan, Mihael, Caoilfhionn Lane, Eoin Ó Droighneáin, and Paul Buitelaar (2016). “IRIS: English-Irish Machine Translation System”. In: *The International Conference on Language Resources and Evaluation.* Portoroz, Slovenia, pp. 566–572.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre (2017). “Learning bilingual word embeddings with (almost) no bilingual data”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada, pp. 451–462.

— (2019). “Unsupervised Neural Machine Translation, a new paradigm solely based on monolingual text.” In: *Procesamiento del Lenguaje Natural* 63.

Aziz, Wilker, Sheila Castilho, and Lucia Specia (2012). “PET: a Tool for Post-editing and Assessing Machine Translation.” In: *The International Conference on Language Resources and Evaluation.* Istanbul, Turkey, pp. 3982–3987.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). “Neural Machine Translation by Jointly Learning to Align and Translate”. In: *CoRR* abs/1409.0473. arXiv: `1409.0473`. URL: `http://arxiv.org/abs/1409.0473`.

Barrault, Loic, Ondřej Bojar, Marta R Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. (2019). “Findings of the 2019 conference on machine translation (wmt19)”. In: *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1).* Florence, Italy, pp. 1–61.

Bayón, María Do Campo and Pilar Sánchez-Gijón (2019). “Evaluating machine translation in a low-resource language combination: Spanish-Galician.” In: *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks.* Dublin, Ireland, pp. 30–35.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico (2016). “Neural versus Phrase-Based Machine Translation Quality: a Case Study”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Austin, Texas: Association for Computational Linguistics, pp. 257–267.

Bentivogli, Luisa, Mauro Cettolo, Marcello Federico, and Federmann Christian (2018). “Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment”. In: *15th International Workshop on Spoken Language Translation 2018.* Bruges, Belgium, pp. 62–69.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri (2016a). "Findings of the 2016 Conference on Machine Translation". In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany: Association for Computational Linguistics, pp. 131–198. DOI: 10.18653/v1/W16-2301. URL: http://www.aclweb.org/anthology/W16-2301.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. (2016b). "Findings of the 2016 conference on Machine Translation". In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany, pp. 131–198.

Bojar, Ondřej, Yvette Graham, Amir Kamran, and Miloš Stanojević (2016c). "Results of the wmt16 metrics shared task". In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 199–231.

Bojar, Ondřej and Aleš Tamchyna (2011). "Improving translation model by monolingual data". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics. Edinburgh, Scotland, pp. 330–336.

Brown, Peter F, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer (1993). "The mathematics of statistical machine translation: Parameter estimation". In: *Computational linguistics* 19.2, pp. 263–311.

Bugliarello, Emanuele and Naoaki Okazaki (July 2020). "Enhancing Machine Translation with Dependency-Aware Self-Attention". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1618–1627. DOI: 10.18653/v1/2020.acl-main.147. URL: https://www.aclweb.org/anthology/2020.acl-main.147.

Cadwell, Patrick, Sheila Castilho, Sharon O'Brien, and Linda Mitchell (2016). "Human factors in machine translation and post-editing among institutional translators". In: *Translation Spaces* 5.2, pp. 222–243.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn (2006). "Re-evaluation the role of BLEU in machine translation research". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy.

Castilho, Sheila, Stephen Doherty, Federico Gaspari, and Joss Moorkens (July 2018). "Approaches to human and machine Translation Quality Assessment". In: Machine Translation: Technologies and Applications 1, pp. 9–38.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio Miceli Barone, and Maria Gialama (2017). "A comparative quality evaluation of PBSMT and NMT using professional translators". In: *Proceedings of the 16th Machine Translation Summit*. Nagoya, Japan, pp. 116–131.

CEF Digital (2020). *What is eTranslation*. URL: `https://ec.europa.eu/cefdigital/wiki/display/CEFDIGITAL/What+is+eTranslation`.

Chen, David and William B Dolan (2011). "Collecting highly parallel data for paraphrase evaluation". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, USA, pp. 190–200.

Chen, Yuan-Lu (2018). "Improving Neural Net Machine Translation Systems with Linguistic Information". PhD thesis. Arizona, USA.

Christian Brothers (1960). *Graiméar Gaeilge na mBráithre Críostaí*. Dublin: M.H. Mac an Ghoill agus a Mhac, Tta.

— (1962). *New Irish Grammar*. Dublin: C J Fallon.

Commons, Creative (2020). *About CC Licenses - Creative Commons*. URL: `https://creativecommons.org/about/cclicenses/`.

Connecting Europe Facility (2020). *CET-AT Service Catalogue.* URL: https://cef-at-service-catalogue.eu/catalogue/browse/e4fc4c58-39fa-43b5-877b-16098ef0d45c/.

Conradh na Gaeilge (2012). *What is Conradh na Gaeilge?* URL: https://cnag.ie/en/info/conradh-na-gaeilge/whatiscnag.html.

Coughlin, Deborah (2003). "Correlating automated and human assessments of machine translation quality". In: *Proceedings of the Ninth Machine Translation Summit*. New Orleans, USA, pp. 63–70.

Defauw, Arne, Sara Szoc, Tom Vanallemeersch, Anna Bardadym, Joris Brabers, Frederic Everaert, Kim Scholte, Koen Van Winckel, and Joachim Van den Bogaert (2019). "Developing a Neural Machine Translation system for Irish". In: *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*. Dublin, Ireland, pp. 32–38.

Denkowski, Michael and Alon Lavie (2012). "Challenges in predicting machine translation utility for human post-editors". In: *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers*. Association for Computational Linguistics. San Diego, USA.

Dowling, Meghan, Lauren Cassidy, Eimear Maguire, Teresa Lynn, Ankit Srivastava, and John Judge (2015). "Tapadóir: Developing a Statistical Machine Translation Engine and Associated Resources for Irish". In: *The 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of the The Fourth LRL Workshop: "Language Technologies in support of Less-Resourced Languages"*. Poznan, Poland, pp. 314–318.

Dowling, Meghan, Sheila Castilho, Joss Moorkens, Teresa Lynn, and Andy Way (2020). "A human evaluation of English-Irish statistical and neural machine translation". In: *Proceedings of the 20th Annual Conference of the European Association for Machine Translation.* On-line (Lisbon), pp. 431–440.

Dowling, Meghan, Teresa Lynn, Yvette Graham, and John Judge (2016). "English to Irish Machine Translation with Automatic Post-Editing". In: *Journées d'Études sur la Parole Traitement Automatique des Langues Naturelles Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (JEP-TALN-RECITAL). The 2nd Celtic Language Technology Workshop*. Paris, France, pp. 42–54.

Dowling, Meghan, Teresa Lynn, Alberto Poncelas, and Andy Way (2018). "SMT versus NMT: Preliminary comparisons for Irish". In: *Technologies for MT of Low Resource Languages*. Boston, USA, pp. 12–20.

Dowling, Meghan, Teresa Lynn, and Andy Way (2017). "A crowd-sourcing approach for translations of minority language user-generated content (UGC)". In: *The 20th Annual Conference of the European Association for Machine Translation. First workshop on Social Media and User Generated Content Machine Translation*, pp. 1–12.

— (2019a). "Investigating backtranslation for the improvement of English-Irish machine translation". In: *TEANGA, the Journal of the Irish Association for Applied Linguistics* 26, pp. 1–25.

— (2019b). "Leveraging backtranslation to improve machine translation for Gaelic languages". In: *Machine Translation Summit XVII*. Vol. 604: *3rd Celtic Language Technology workshop*. Dublin, Ireland, pp. 58–62.

Dryer, Matthew S. and Martin Haspelmath, eds. (2013). *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: `https://wals.info/`.

Duh, Kevin, Paul McNamee, Matt Post, and Brian Thompson (2020). "Benchmarking neural and statistical machine translation on low-resource african languages". In: *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 2667–2675.

Edunov, Sergey, Myle Ott, Michael Auli, and David Grangier (2018). "Understanding Back-Translation at Scale". In: *Proceedings of the 2018 Conference on Em-*

pirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, pp. 489–500.

Escartín, Carla Parra and Manuel Arcedillo (Oct. 2015). "Living on the edge: productivity gain thresholds in machine translation evaluation metrics". In: *Proceedings of the 4th Workshop on Post-editing Technology and Practice*. Miami, USA, p. 46.

Esplá-Gomis, Miquel and Mikel Forcada (2009). "Bitextor, a free/open-source software to harvest translation memories from multilingual websites". In: *Proceedings of MT Summit XII*. Ottawa, Canada: Association for Machine Translation in the Americas.

Esplà-Gomis, Miquel, Mikel L Forcada, Gema Ramírez-Sánchez, and Hieu Hoang (2019). "ParaCrawl: Web-scale parallel corpora for the languages of the EU". In: *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*. Dublin, Ireland, pp. 118–119.

Etchegoyhen, Thierry, Borja Anza Porras, Andoni Azpeitia, Eva Martínez Garcia, Paulo Vale, José Luis Fonseca, Teresa Lynn, Jane Dunne, Federico Gaspari, Andy Way, et al. (2018). "ELRI. European Language Resource Infrastructure". In: *The 21st Annual Conference of the European Association for Machine Translation*. Alacant, Spain: European Association for Machine Translation, p. 351.

Etchegoyhen, Thierry, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard van Loenhout, Arantza del Pozo, Mirjam Sepesy Maucec, Anja Turner, and Martin Volk (2014). "Machine translation for subtitling: a large-scale evaluation". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Reykjavik, Iceland.

European Data Portal (2020). URL: `https://www.europeandataportal.eu/en/news/european-language-resource-coordination`.

Fadaee, Marzieh and Christof Monz (2018). "Back-Translation Sampling by Targeting Difficult Words in Neural Machine Translation". In: *Proceedings of the 2018*

Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, pp. 436–446.

Forcada, Mikel L (2017). “Making sense of neural machine translation”. In: *Translation spaces* 6.2, pp. 291–309.

Forcada, Mikel L, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O’Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers (2011). “Apertium: a free/open-source platform for rule-based machine translation”. In: *Machine translation* 25.2, pp. 127–144.

Gage, Philip (1994). “A new algorithm for data compression”. In: *The C Users Journal* 12.2, pp. 23–38.

Gain, Vish (2021). *Funding for two AI projects to prevent the ‘digital extinction’ of Irish*. URL: `https://www.siliconrepublic.com/machines/irish-language-digital-research-funding-machine-translation`.

Galley, Michel and Christopher D Manning (2008). “A simple and effective hierarchical phrase reordering model”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. Honolulu, USA, pp. 848–856.

Gong, Zhengxian, Min Zhang, and Guodong Zhou (2011). “Cache-based document-level statistical machine translation”. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, United Kingdom, pp. 909–919.

Graham, Yvette, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi (2016). “Is all that glitters in machine translation quality estimation really gold?” In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 3124–3134.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel (2013). “Continuous measurement scales in human evaluation of machine translation”. In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Osaka, Japan, pp. 33–41.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel (2017). "Can machine translation systems be evaluated by the crowd alone". In: *Natural Language Engineering* 23.1, pp. 3–30.

Guerberof Arenas, Ana (2008). "Productivity and quality in the post-editing of outputs from translation memories and machine translation". In: *Localisation Focus* 7.1, pp. 11–21.

Guzmán, Francisco, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato (Nov. 2019). "The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali–English and Sinhala–English". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 6098–6111. DOI: 10.18653/v1/D19-1632. URL: https://aclanthology.org/D19-1632.

Heafield, Kenneth (2011). "KenLM: Faster and smaller language model queries". In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland, pp. 187–197.

Hearne, Mary and Andy Way (2011). "Statistical machine translation: a guide for linguists and translators". In: *Language and Linguistics Compass* 5.5, pp. 205–226.

Higgins, Noelle and Dorothy Ní Uigín (2017). "Irish language in the courts: Is there a need for, and a right to, an interpreter?" In: *Legal translation and court interpreting: ethical values, quality, competence training* 140, p. 49.

Hoang, Vu Cong Duy, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn (2018). "Iterative back-translation for neural machine translation". In: *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*. Melbourne, Australia, pp. 18–24.

Imankulova, Aizhan, Raj Dabre, Atsushi Fujita, and Kenji Imamura (2019). "Exploiting Out-of-Domain Parallel Data through Multilingual Transfer Learning for

Low-Resource Neural Machine Translation". In: *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*. Dublin, Ireland, pp. 128–139.

Jassem, Krzysztof and Tomasz Dwojak (2019). "Statistical versus neural machine translation–a case study for a medium size domain-specific bilingual corpus". In: *Poznan Studies in Contemporary Linguistics* 55.2, pp. 491–515.

Judge, John, Ailbhe Ní Chasaide, Rose Ní Dhubhda, Kevin P. Scannell, and Elaine Uí Dhonnchadha (2012). *An Ghaeilge sa Ré Dhigiteach – The Irish Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Available online at `http://www.meta-net.eu/whitepapers`. Springer. ISBN: 978-3-642-30557-3.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. URL: `http://arxiv.org/abs/1412.6980`.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush (2017). "OpenNMT: Open-Source Toolkit for Neural Machine Translation". In: *Proceedings of ACL 2017, System Demonstrations*. Vancouver, Canada, pp. 67–72.

Kocmi, Tom and Ondřej Bojar (2018). "Trivial Transfer Learning for Low-Resource Neural Machine Translation". In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Brussels, Belgium, pp. 244–252.

Koehn, Philipp (2009). *Statistical machine translation*. New York, USA: Cambridge University Press.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, and Richard Zens (2007). "Moses: Open source toolkit for statistical machine translation". In: *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Prague, Czech Republic, pp. 177–180.

Koehn, Philipp and Rebecca Knowles (2017). “Six Challenges for Neural Machine Translation”. In: *Proceedings of the First Workshop on Neural Machine Translation*. Vancouver, Canada, pp. 28–39.

Koehn, Philipp and Christof Monz (June 2006). “Manual and Automatic Evaluation of Machine Translation between European Languages”. In: *Proceedings on the Workshop on Statistical Machine Translation*. New York City, USA: Association for Computational Linguistics, pp. 102–121. URL: https://aclanthology.org/W06-3114.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu (May 2003). “Statistical Phrase-based Translation”. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HT), Volume 1*. Edmonton, Canada, pp. 48–54.

Lakew, Surafel M, Mattia A Di Gangi, and Marcello Federico (2017). “Multilingual Neural Machine Translation for Low Resource Languages”. In: *Fourth Italian Conference on Computational Linguistics*. Rome, Italy, p. 189.

Lohar, Pintu (2020). “Machine translation of user-generated content”. PhD thesis. Dublin City University.

Lommel, Arle, Hans Uszkoreit, and Aljoscha Burchardt (2014). “Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics”. In: *Tradumàtica* 12, pp. 0455–463.

Luong, Thang, Hieu Pham, and Christopher D Manning (2015). “Effective Approaches to Attention-based Neural Machine Translation”. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal, pp. 1412–1421.

Lynn, Teresa, Micheál Ó Conaire, and Jane Dunne (2019). “Country Profile Ireland”. In: *ELRC WHITE PAPER: Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe*. Saarbrucken: German Research Center for Artificial Intelligence (DFKI), pp. 92–97. ISBN: 978-3-943853-05-6.

Lynn, Teresa and Kevin Scannell (2019). “Code-switching in Irish tweets: A preliminary analysis”. In: *Proceedings of the Celtic Language Technology Workshop*. Dublin, Ireland, pp. 32–40.

Martindale, Marianna, Marine Carpuat, Kevin Duh, and Paul McNamee (2019). “Identifying fluently inadequate output in neural and statistical machine translation”. In: *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pp. 233–243.

Mendels, Gideon, Victor Soto, Aaron Jaech, and Julia Hirschberg (May 2018). “Collecting Code-Switched Data from Social Media”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: European Language Resources Association (ELRA). ISBN: 979-10-95546-00-9.

Miculicich, Lesly, Dhananjay Ram, Nikolaos Pappas, and James Henderson (Oct. 2018). “Document-Level Neural Machine Translation with Hierarchical Attention Networks”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2947–2954. DOI: 10.18653/v1/D18-1325. URL: https://aclanthology.org/D18-1325.

Moorkens, Joss (2020). “Comparative satisfaction among freelance and directly-employed Irish-language translators”. In: *The International Journal for Translation & Interpreting Research* 12.1, pp. 55–73. DOI: 10.12807/ti.112201.2020.a04. URL: http://www.trans-int.org/index.php/transint/article/view/801.

Moorkens, Joss and Sharon O’Brien (2015). “Post-editing evaluations: Trade-offs between novice and professional participants”. In: *Proceedings of the 18th An-*

nual Conference of the European Association for Machine Translation. Antalya, Turkey, pp. 75–81.

Moorkens, Joss, Antonio Toral, Sheila Castilho, and Andy Way (2018). “Translators’ perceptions of literary post-editing using statistical and neural machine translation”. In: *Translation Spaces* 7.2, pp. 240–262.

Murray, Kenton, Jeffery Kinnison, Toan Q Nguyen, Walter Scheirer, and David Chiang (2019). “Auto-Sizing the Transformer Network: Improving Speed, Efficiency, and Performance for Low-Resource Machine Translation”. In: *Proceedings of the 3rd Workshop on Neural Generation and Translation*. Hong Kong, pp. 231–240.

Mutal, Jonathan David, Lise Volkart, Pierrette Bouillon, Sabrina Girletti, and Paula Susana Estrella (2019). “Differences between SMT and NMT Output-a Translators’ Point of View”. In: *The Second Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*. Varna, Bulgaria.

Neubig, Graham, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang (2019). “compare-mt: A Tool for Holistic Comparison of Language Generation Systems”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota, pp. 35–41.

Ó Maolalaigh, Roibéard (2016). “DASG: Digital Archive of Scottish Gaelic/Dachaigh airson Stòras na Gàidhlig”. In: *Scottish Gaelic Studies* 30, pp. 242–262.

Ó Múrchú, Pól (2013). *A Grammar of Modern Irish: An Annotated Guide to Graiméar Gaeilge na mBráithre Críostaí*. Baile Átha Cliath.

Ó Siadhail, Mícheál (1989). *Modern Irish: Grammatical structure and dialectal variation*. Cambridge University Press.

O’Brien, Sharon, Michel Simard, and Marie-Josée Goulet (2018). “Machine translation and self-post-editing for academic writing support: Quality explorations”. In: *Translation quality assessment*. Springer, pp. 237–262.

Och, Franz Josef, Nicola Ueffing, and Hermann Ney (2001). “An Efficient A* Search Algorithm for Statistical Machine Translation”. In: *Proceedings of the Workshop* *on Data-Driven Methods in Machine Translation, Volume 14*. Toulouse, France, pp. 1–8.

Ondřej, Bojar, Rajen Chatterjee, Federmann Christian, Graham Yvette, Haddow Barry, Huck Matthias, Koehn Philipp, Liu Qun, Logacheva Varvara, Monz Christof, et al. (2017). "Findings of the 2017 conference on machine translation (wmt17)". In: *Second Conference on Machine Translation*. Copenhagen, Denmark, pp. 169–214.

Papavassiliou, Vassilis, Prokopis Prokopidis, and Gregor Thurmair (Aug. 2013). "A modular open-source focused crawler for mining monolingual and bilingual corpora from the web". In: *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 43–51. URL: http://www.aclweb.org/anthology/W13-2506.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Philadelphia, Pennsylvania, pp. 311–318.

Passban, Peyman, Qun Liu, and Andy Way (2017). "Translating Low-Resource Languages by Vocabulary Adaptation from Close Counterparts". In: *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 16.4, p. 29.

Petukhova, Volha, Rodrigo Agerri, Mark Fishel, Sergio Penkale, Arantza del Pozo, Mirjam Sepesy Maučec, Andy Way, Panayota Georgakopoulou, and Martin Volk (May 2012). "SUMAT: Data Collection and Parallel Corpus Compilation for Machine Translation of Subtitles". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 21–28. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/154_Paper.pdf.

Poncelas, Alberto, Kepa Sarasola, Meghan Dowling, Andy Way, Gorka Labaka, and Iñaki Alegria (2019). "Adapting NMT to caption translation in Wikimedia

Commons for low-resource languages". In: *Procesamiento del Lenguaje Natural* 63, pp. 33–40.

Poncelas, Alberto, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban (2018). "Investigating backtranslation in neural machine translation". In: *Proceedings of the 21st Annual Conference of the European Association for Machine Translation.* Alacant, Spain, pp. 249–258.

Popović, Maja (2015). "chrF: character n-gram F-score for automatic MT evaluation". In: *Proceedings of the Tenth Workshop on Statistical Machine Translation.* Lisbon, Portugal, pp. 392–395.

— (2017). "Comparing language related issues for NMT and PBMT between German and English". In: *The Prague Bulletin of Mathematical Linguistics* 108.1, p. 209.

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

Prys, Delyth and Dewi Bryn Jones (2018). "Gathering data for speech technology in the welsh language: A case study". In: *Proceedings of the LREC 2018 Workshop "CCURL*. Miyazaki, Japan.

Publications Office of the European Union (2011). Publications Office of the European Union.

Raghallaigh, Brian Ó, Kevin Scannell, and Meghan Dowling (2019). "Improving full-text search results on duchas.ie using language technology". In: *Machine Translation Summit XVII. Proceedings of the Celtic Language Technology Workshop*. Dublin, Ireland, pp. 63–69.

Rannóg an Aistriucháin (1962). *Gramadach na Gaeilge agus Litriú na Gaeilge: An Caighdeán Oifigiúil.* Baile Átha Cliath: Oifig an tSoláthair.

Reiter, Ehud (2018). "A Structured Review of the Validity of BLEU". In: *Computational Linguistics* 44.3, pp. 393–401.

Riagáin, Ó' (2017). "Irish". In: *Language in the British Isles*, pp. 218–236.

Rubino, Raphael, Tommi A Pirinen, Miquel Espla-Gomis, Nikola Ljubešić, Sergio Ortiz Rojas, Vassilis Papavassiliou, Prokopis Prokopidis, and Antonio Toral (2015). “Abu-matran at wmt 2015 translation task: Morphological segmentation and web crawling”. In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Lisbon, Portugal, pp. 184–191.

Saldanha, Gabriela and Sharon O’Brien (2014). *Research methodologies in translation studies*. Routledge.

Scannell, Kevin (2006). “Machine translation for closely related language pairs”. In: *Proceedings of the Workshop Strategies for developing machine translation for minority languages*. Citeseer. Genoa, Italy, pp. 103–109.

— (2014). “Statistical models for text normalization and machine translation”. In: *Proceedings of the First Celtic Language Technology Workshop*. Dublin, Ireland, pp. 33–40.

Sennrich, Rico and Barry Haddow (2016). “Linguistic Input Features Improve Neural Machine Translation”. In: *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*. Vol. 1. Berlin, Germany, pp. 83–91.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016a). “Improving Neural Machine Translation Models with Monolingual Data”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. Berlin, Germany, pp. 86–96.

— (2016b). “Neural Machine Translation of Rare Words with Subword Units”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. Berlin, Germany, pp. 1715–1725.

Sennrich, Rico and Biao Zhang (2019). “Revisiting Low-Resource Neural Machine Translation: A Case Study”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, pp. 211–221.

Shterionov, Dimitar, Félix do Carmo, Joss Moorkens, Murhaf Hossari, Joachim Wagner, Eric Paquin, Dag Schmidtke, Declan Groves, and Andy Way (2020). “A roadmap to neural automatic post-editing: an empirical approach". In: *Machine Translation* 34.2, pp. 67–96.

Shterionov, Dimitar, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'Dowd, and Andy Way (2018). "Human versus automatic quality evaluation of NMT and PBSMT". In: *Machine Translation* 32.3, pp. 217–235.

ShweSin, Yi Mon, Khin Mar Soe, and Khin Yadanar Htwe (2018). "Large scale Myanmar to English neural machine translation system". In: *2018 IEEE 7th Global Conference on Consumer Electronics (GCCE)*. IEEE. Las Vegas, USA, pp. 464–465.

Siddhant, Aditya, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu (July 2020). "Leveraging Monolingual Data with Self-Supervision for Multilingual Neural Machine Translation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 2827–2835. DOI: 10.18653/v1/2020.acl-main.252. URL: https://www.aclweb.org/anthology/2020.acl-main.252.

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

Snover, Matthew G, Nitin Madnani, Bonnie Dorr, and Richard Schwartz (2009). "TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate". In: *Machine Translation* 23.2-3, pp. 117–127.

Soto, Xabier, Dimitar Shterionov, Alberto Poncelas, and Andy Way (July 2020). "Selecting Backtranslated Data from Multiple Sources for Improved Neural Machine Translation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 3898–3908. URL: https://www.aclweb.org/anthology/2020.acl-main.359.

The European Commission (2019). *Report From the Commission to the Council on the Union institutions' progress towards the implementation of the gradual reduction of the Irish language derogation.* Brussels: The European Commission. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52019DC0318&from=en.

— (2020). *The Irish language in the EU: on the way to full status.* URL: https://ec.europa.eu/ireland/news/key-eu-policy-areas/the-irish-language-in-the-eu_en.

Tiedemann, Jorg (May 2012a). "Parallel Data, Tools and Interfaces in OPUS". In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis. Istanbul, Turkey: European Language Resources Association (ELRA). ISBN: 978-2-9517408-7-7.

— (2012b). "Parallel Data, Tools and Interfaces in OPUS." In: *LREC*. Vol. 2012, pp. 2214–2218.

Toral, Antonio (2019). "Post-editese: an Exacerbated Translationese". In: *Proceedings of Machine Translation Summit XVII Volume 1: Research Track.* Dublin, Ireland, pp. 273–281.

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way (2018). "Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation". In: *Proceedings of the Third Conference on Machine Translation: Research Papers.* Brussels, Belgium, pp. 113–123.

Toral, Antonio, Miquel Esplá-Gomis, Filip Klubička, Nikola Ljubešić, Vassilis Papavassiliou, Prokopis Prokopidis, Raphael Rubino, and Andy Way (2017). "Crawl and crowd to bring machine translation to under-resourced languages". In: *Language resources and evaluation* 51.4, pp. 1019–1051.

Tu, Zhaopeng, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li (2016). "Modeling Coverage for Neural Machine Translation". In: *Proceedings of the 54th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1. Berlin, Germany, pp. 76–85.

Vanmassenhove, Eva, Dimitar Shterionov, and Andy Way (2019). "Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation". In: *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*. Dublin, Ireland, pp. 222–232.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *Advances in neural information processing systems*. California, USA, pp. 5998–6008.

Vogel, Stephan, Hermann Ney, and Christoph Tillmann (1996). "HMM-based word alignment in statistical translation". In: *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*. Copenhagen, Denmark.

Wang, Weiyue, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney (2016). "Character: Translation edit rate on character level". In: *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Berlin, Germany, pp. 505–510.

Wang, Xing, Zhaopeng Tu, and Min Zhang (2018). "Incorporating statistical machine translation word knowledge into neural machine translation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.12, pp. 2255–2266.

Way, Andy (2013). "Traditional and emerging use-cases for machine translation". In: *Proceedings of Translating and the Computer* 35, pp. 1–12.

— (2019). "Machine translation: where are we at today?" In: *The Bloomsbury Companion to language industry studies*. Bloomsbury Publishing, pp. 311–332.

Way, Andy and Federico Gaspari (2019). "PRINCIPLE: Providing Resources in Irish, Norwegian, Croatian and Icelandic for the Purposes of Language Engineering". In: *The 3rd Celtic Language Technology Workshop*. Dublin, Ireland, pp. 112–113.

Wenzek, Guillaume, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and édouard Grave (2020). "CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data". In: *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France, pp. 4003–4012.

Werlen, Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson (2018). "Document-Level Neural Machine Translation with Hierarchical Attention Networks". In: *CoRR* abs/1809.01576. arXiv: 1809.01576.

Wieting, John, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig (2019). "Beyond BLEU: Training Neural Machine Translation with Semantic Similarity". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy, pp. 4344–4355.

World Intellectual Property Organization (2021). *Berne Convention Accession by Ireland*. URL: https://www.wipo.int/treaties/en/notifications/berne/treaty_berne_246.html.

Zens, Richard, Franz Josef Och, and Hermann Ney (July 2002). "Phrase-Based Statistical Machine Translation". In: *Proceedings of the Annual Conference on Artificial Intelligence (AAAI)*. Edmonton, Canada, pp. 18–32.

Zhang, Biao, Deyi Xiong, Jun Xie, and Jinsong Su (2020). "Neural Machine Translation With GRU-Gated Attention Model". In: *IEEE Transactions on Neural Networks and Learning Systems* 31.11, pp. 4688–4698. DOI: 10.1109/TNNLS.2019.2957276.

Zhou, Jiawei and Phillip Keung (July 2020). "Improving Non-autoregressive Neural Machine Translation with Monolingual Data". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1893–1898. DOI: 10.18653/v1/2020.acl-main.171. URL: https://www.aclweb.org/anthology/2020.acl-main.171.

Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight (2016). "Transfer Learning for Low-Resource Neural Machine Translation". In: *Proceedings of the*

2016 Conference on Empirical Methods in Natural Language Processing. Texas, USA, pp. 1568–1575.