Gaeilge Bhriste ó Shamhlacha Cliste: How Clever Are LLMs When Translating Irish Text?

Teresa Clifford
Fiontar & Scoil na Gaeilge
Dublin City University
ADAPT Centre
Dublin City University
teresa.clifford3@mail.dcu.ie

Abigail Walsh and Brian Davis
ADAPT Centre
Dublin City University
abigail.walsh@adaptcentre.ie
Brian.Davis@adaptcentre.ie

Micheál J. Ó Meachair
Fiontar & Scoil na Gaeilge
Dublin City University
micheal.omeachair@dcu.ie

Abstract

Large Language Models have been widely adopted in NLP tasks and applications, however, their ability to accurately process Irish and other minority languages has not been fully explored. In this paper we describe preliminary experiments examining the capacity of publicly-available machine translation engines (Google Translate, Microsoft Bing, and eTranslation) and prompt-based AI systems systems (ChatGPT 3.5, Llama 2) for translating and handling challenging language features of Irish. A hand-crafted selection of challenging Irish language features were incorporated into translation prompts, and the output from each model was examined by a human evaluator. The results of these experiments indicate that these LLM-based models still struggle with translating rare linguistic phenomena and ambiguous constructions. This preliminary analysis helps to inform further research in this field, providing a simple ranking of publicly-available models, and indicating which language features require particular attention when evaluating model capacity.

1 Introduction

The rising interest in transformer-based Large Language Models (LLMs) in the field of Natural Language Processing (NLP) can be seen in the high volume of publications continually being published in major computational linguistics venues year by year (e.g. LREC: (Ekgren et al., 2022); ACL: (Raunak et al., 2023), and (Wu and Hu, 2023); EACL: (Balloccu et al., 2024)), as well as increased use of ChatGPT and similar applications in people's daily lives. As hype surrounding these models continues to build with improvements in performance, the question arises of how the field of machine translation is impacted, and whether machine translation can be considered a 'solved problem' (Zhu et al., 2024).

Despite ongoing discussion, the field lacks depth of understanding on the ability of these models to process minority languages, including Irish. This paper describes preliminary experiments in order to shed light on the ability of publicly-available machine translation (MT) engines and prompt-based AI systems when translating certain hand-selected challenging features of the Irish language (e.g. non-compositional constructions such as *Bóín Dé* (God's little cow) 'ladybird').

Relevant background and related work is explored in Section 2, and Section 3 describes the experimental set up. The results of the experiments are recorded in Section 4, and include a human evaluation of the target translations. The experiments represent the initial steps in a thorough exploration of the capacity of LLM-based systems to process text from low-resourced languages such as Irish. Section 6 explores future areas for exploration in this research topic.

2 Background

2.1 Machine Translation for Irish

Irish is the official language of Ireland and an official EU language. Despite this status, the language is considered a low-resource language by European language researchers (Lynn, 2022), noted to have weak or no support in many categories of technological support for selected European languages, similar to West Frisian (Robinson-Jones and Scarse, 2022) and other minority languages. Lynn (2022) discusses how subpar applications and language

Proceedings of the 5th Celtic Language Technology Workshop (CLTW 5), pages 46–51
Abu Dhabi, UAE (Online), January 20, 2025

tools are a factor that can lead to Irish speakers switching to using English in online spaces, which contributes to the rising risk of digital extinction for the Irish language. To address this threat, the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media launched The Digital Plan for the Irish Language in December of 2022 (Ní Chasaide et al.). This plan calls attention to several areas of research that are vital to the advancement of Irish language technology, including Machine Translation (MT) and the development of key resources.

MT is an area of Irish language technology that has seen slow but relatively consistent development over past years, with publications demonstrating recent advances in the field, e.g. applying cutting edge methodology to building bidirectional English & Irish (EN<>GA) MT models (Lankford et al., 2024), and focusing on domain-specific translation (Lankford et al., 2021).

Irish is one of the languages supported by eTranslation (European Commission), an open-source MT platform developed in partnership with the European Commission. General-purpose MT systems, such as Google Translate (Google Research) and Microsoft Bing Translator (Microsoft Research), also offer support for the Irish language. Research group ABAIR (ABAIR) have developed applications with chatbot-style interactions primarily for Computer Assisted Language Learning (CALL) and grammar checking purposes; focusing on speech-to-text and text-to-speech technology (e.g. *An Scéalaí*[1], *An Bat Mírialta*[2]). However, it is difficult to assess the true capacity of many prompt-based AI systems to correctly handle Irish text. Some multi-lingual models (e.g. Gliglish)[3] may claim support for Irish language, but omit details on how the model has been trained, and what data was included in training. For example, when tested with Irish prompts, the Gliglish model showed substantial problems in the output, providing nonsensical replies (e.g. *Go raibh maith agat, táim ag cuir blasta ort agus beidh mé ag dul le... arán* meaning "Thank you, I am putting tasty on you and I will be going with... bread").

2.2 Translation Difficulty

Translation difficulty is often described in terms of human translators and their mental or cognitive load (Akbari and Segers, 2017; Sun, 2015). However, there is an overlap between the translation difficulty for human translators and MT systems (Vanroy et al., 2019). O'Brien (2004) examined the effect of **N**egative **T**ranslatability **I**ndicators (NTIs)—i.e. linguistic features that have been noted as problematic for MT (Gdaniec, 1994; Bernth, 1999; Bernth and Gdaneic, 2001; Underwood and Jongejan, 2001), such as the passive voice, and the gerund—on the post editing effort. The data suggested that the post-editing speed for sentences without NTIs was faster than those with them on average. Some of these NTIs, such as lexical ambiguity, also fall under the umbrella of translation ambiguity (Tokowicz, 2014). Examples of translation ambiguity can be seen in our tested language features (e.g. lexical ambiguity; one word having two meanings in one language).

3 Methodology

Experiments were set up to test the capacity of publicly-available MT engines and prompt-based AI models on translating certain hand-select NTIs, incorporated into translation examples and translation prompts in either English or Irish, based on the feature used for evaluation. Two rounds of experiments took place, with different translation examples selected for each round.

Six challenging features of the Irish language were selected for testing in Round One, with four additional features tested in Round Two. These features were chosen based on previous work on challenging features of Irish language (e.g. Walsh et al. (2019)), in research on translation difficulty in other languages (Tokowicz, 2014), and based on the the researchers' knowledge of the Irish language.

The features chosen for Round One were:

1. Words that have multiple meanings (e.g. homonyms "bark" the sound made by dogs vs "bark" the protective covering on trees)

2. Words that do not have direct translations in one language (e.g. *Súilaithne* means 'to know someone to see')

3. Non-compositional phrases, where the combined meanings of the individual words in a phrase are not equivalent to the meaning of the phrase (e.g. *Duilleog bháite* (drowned leaf) 'water lily')

4. Phrases including 'yes' and 'no', as there is no direct translation for these words in Irish

[1] https://scealai.abair.ie/

[2] https://bat-mirialta.abair.ie/

[3] https://gliglish.com/

5. Phrases using the construction ‘I am’, as there are two verbs for ‘be’ in Irish: copular and substantive ‘be’

6. Uncensored swear words and innuendo (e.g. ‘I fucked her’)

The additional features included in Round Two were as follows:

7. *Logainmneacha* or Irish place names (e.g. *Baile Átha Cliath* ‘Dublin’)

8. *An tuiseal gairmneach* or the vocative case (e.g. *A **Sheáin*** features slenderisation and lenition in vocative case)

9. Non-compositional animal names (e.g. *Mac tíre* (son of the land) ‘wolf’)

10. Mythical creature names (e.g. *Bean Sí* (fairy woman) ‘banshee’)

The models chosen for Round One of these experiments were Google Translate (Google Research), Microsoft Bing Translator (Microsoft Research), eTranslation (European Commission) and ChatGPT 3.5 (OpenAI, 2024), with Llama 2 (Meta AI) being additionally included for the Round Two.[4] These applications were chosen as they are all publicly available, free to use,[5] and state that they can translate from English to Irish and Irish to English. This was assessed by the inclusion of Irish as one of the language options on the language list for MT applications, or by prompting the AI system, asking if it has the capability to translate English to Irish and vice versa. Prompt-based AI systems Gliglish and Gemini were originally considered for inclusion but were rejected due to their use cases not fitting the experiment parameters, with 1) Gliglish only accepting speech input, and 2) Gemini expressing it had the ability to translate to and from Irish when initially prompted in English, then stating it was unable to do so when asked directly to translate words or sentences provided in Irish. The applications were tested using default settings, with no changes to add advanced search features where these features were offered by the application.

Examples were hand-crafted words or sentences in Irish or English, integrating one of the listed features. New examples were crafted for Round Two, which integrated the additional language features and also new examples of the language features from Round One, often adjusted to include more specific context words, as informed by the research of Castilho and Knowles (2024) and Castilho et al. (2020) (e.g. Round One example: *Chonaic mé bóín Dé thíos ansin.* ‘I saw a ladybird down there’ vs. Round Two example: *Is feithid é bóín Dé* ‘A lady bird is an insect’). Round One contained 57 examples, and Round Two contained 132 examples, for a total of 189 examples. Each example was manually fed into each system interface, and the outputs were recorded.

When collecting the translation outputs, only the first translation provided by each model was recorded, even when alternative translations were offered. Both ChatGPT 3.5 and Llama 2 were given the following initial prompt before the examples were provided: “Hello can you translate these sentences and words from English to Irish or from Irish to English please”. Any extra context or information provided by the AI systems was also recorded.

4 Results

Given the range of potentially correct translations, a manual evaluation was deemed a more reliable means of capturing the models’ capabilities rather than automatic metrics, e.g. BLEU. An assessment was made by a fluent Irish speaker to determine whether a target translation was ‘plausible’, i.e. a translation that may be incorrect due to the context of the example (i.e a direct translation of a non-compositional phrase) but there could be cases in which this translation is correct in context. ‘High-quality’ translations were those considered a correct and adequate translation, without grammatical error, and correct in context to the example. Table 1 displays the results of this assessment for the examples produced by each system, summed from Round One and Round Two. In Figure 1, the percentage of ‘plausible’ and ‘high-quality’ translations produced by each system are calculated for each feature, to indicate the general level of challenge each feature presents.

F3, F9, F10 exhibit the largest divide between ‘plausible’ and ‘high-quality’ translations. This suggests the systems are attempting to translate words

[4] Models used were the most up-to-date version of the prompt-based AI systems at the time of the experiments.

[5] It should be noted that, while free to use, an account must be created to use eTranslation and ChatGPT.

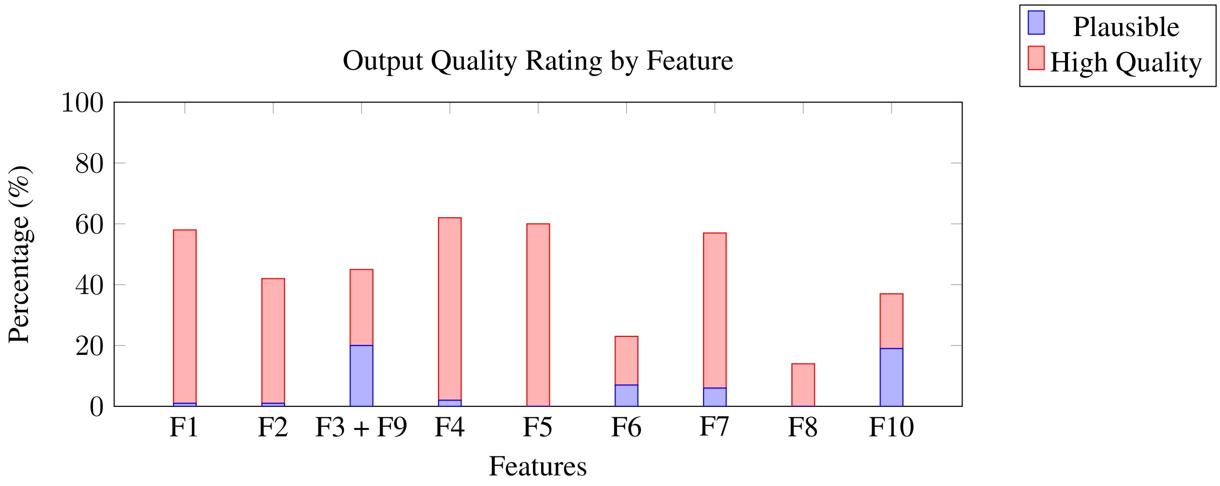

Figure 1: Percentage of 'plausible' and 'high-quality' translations per language feature. F1 stands for Feature 1, referring to the first of the features listed in Section 3. Non-compositional phrases (F3) and Non-compositional animal (F9) names were combined into F3 + F9.

System	Plausible	High-quality
Bing	51.3%	36.5%
Google	47.6%	38.1%
eTranslation	45%	32.3%
ChatGPT	46%	38.1%
Llama	28%	16%

Table 1: Percentage of 'plausible' and 'high-quality' translations for each system (rounded to the nearest decimal point) for both Round One and Round Two, out of a total of 189. Llama was used to test 132 examples, as it was not included in Round One.

that are out-of-vocabulary or rarely represented in the training data, leading to producing literal or word-for-word translations which could be classified as 'plausible' but not 'high-quality'. The issue of rare or out-of-vocabulary words may also be the case for F6, as swear words are likely filtered out of the training data. Similarly, with F8 and F10, mythical creatures and the vocative case may not be heavily represented in training data.

100% of the 'plausible' translations for features F5 and F8 were also 'high quality', which is intuitive as translations for these features can only be correct or incorrect. However, the rate of plausibility by systems for these features was only 60% and 14% respectively, indicating that systems struggled in particular with correct handling of the vocative case.

Table 2, provides the rate of plausibility achieved by each system over the different rounds of the experiments, to provide an overview of the system's capabilities to translate these features as a whole.

System	Round One	Round 2	Both Rounds
Bing	51%	54%	52.5%
Google	47%	48%	47.5%
eTranslation	51%	42%	46.5%
ChatGPT	49%	45%	47%
Llama		28%	28%

Table 2: Rate of plausibility achieved by the systems in Round One, Round Two and across both rounds (rounded to the nearest decimal point).

Microsoft Bing Translator was the most successful model for producing plausible translations of these challenging language features. Llama 2 was the least successful model overall. Of the models that were tested in both Rounds of the experiment, the eTranslation model was slightly less successful than the ChatGPT and Google models.

5 Conclusions

Despite having the highest rate of plausibility, Microsoft Bing Translator had an almost 50% rate of implausible translations. Even the features whose 'plausible' translations were all also 'high-quality', had rates of plausibility as low as 14%. From these initial experiments, it appears that LLMs and publicly-available MT models are currently not adequately supported for these challenging features of the Irish language, particularly for rare linguistic words and features, such as the vocative case, and swear words.

6 Future Work

Future experiments will aim to automate the input and prompting phase of the experiment, in order to increase the size of the test data. We also aim to include additional models (e.g. bespoke Irish encoder-decoder models, or other publicly-available models that support use of the Irish language). Additionally, we aim to expand the number of challenging language features explored; such as including culturally distinct words and phrases (e.g. 'foot path' in Ireland vs 'side walk' in the USA). Future experiments will include baseline examples, where each challenging feature is substituted with a non-challenging feature, in order to compare the capability of each model to translate a non-challenging example of the same syntactic or lexical form. Potential categorisation of the challenging features would help with this step (e.g. grouping lexically challenging examples, grammatically challenging examples, ambiguous examples), which will further inform the capacity of each model to handle different types of challenging language. Other experiment adjustments include prompting the AI systems systems in Irish as opposed to English.

7 Limitations

This research represents a preliminary study, exploring the results of including a small hand-crafted selection of examples of difficult-to-translate features of Irish.

A researcher with Irish language skills equivalent to a C2 level[6] developed the test set, and performed the analysis of the results. This limits the scope of the analysis. Words and phrases can have a variety of different meanings, and a single person cannot capture this variety. Not only could multiple researchers increase the likelihood of noticing any mistakes in typos in the test set, they would also hep ensure that valid translations that differ from one researcher's preferred translation would be captured. This would be particularly useful in the context of the Irish language, as a native speaker's dialect may influence what they would see as a correct translation.

These limitations acknowledged, these experiments provide an initial comparison of systems for automatic translation, indicates particularly problematic features that require more investigation, and leaves room for future experiments incorporating these insights and adjusted methodology.

[6] According to CEFR Levels provided here: https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale

Acknowledgments

eSTÓR's ongoing operations and growth are sustained through funding from the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media. We would also like to thank the reviewers for their detailed feedback, many of whose comments were incorporated into the final paper.

References

ABAIR. Abair.ie. https://abair.ie/ga.

Alireza Akbari and Winibert Segers. 2017. Translation Difficulty: How to Measure and What to Measure. *Lebende Sprachen*, pages pp. 3–29.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Arendse Bernth. 1999. A confidence index for Machine Translation. In *Proceedings of the 8th Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, University College, Chester.

Arendse Bernth and Claudia Gdaneic. 2001. MTranslatabilityy. *Machine Translation, Volume 16, Issue 3*, pages pp. 175–218.

Sheila Castilho and Rebecca Knowles. 2024. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, pages pp. 1–31.

Sheila Castilho, Maja Popović, and Andy Way. 2020. On Context Span Needed for Machine Translation Evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages pp. 3735–3742, Marseille, France. European Language Resources Association.

Ariel Ekgren, Amaru Cuba Gyllensten, Evangelia Gogoulou, Alice Heiman, Severine Verlinden, Joey Öhman, Fredrik Carlsson, and Magnus Sahlgren. 2022. Lessons Learned from GPT-SW3: Building the First Large-Scale Generative Language Model for Swedish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3509–3518, Marseille, France. European Language Resources Association.

European Commission. eTranslation - The European Commission's Machine Translation System. https://commission.europa.eu/resources-partners/etranslation_en.

Claudia Gdaniec. 1994. The Logos Translatability Index. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, USA.

Google Research. Google Translate. https://translate.google.com/?sl=en&tl=ga&op=translate.

Seamus Lankford, Haithem Afli, and Andy Way. 2021. Machine Translation in the Covid domain: an English-Irish case study for LoResMT 2021. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)s*, pages 144–150, Virtual. Association for Machine Translation in the Americas.

Séamus Lankford, Haithem Afli, and Andy Way. 2024. Design of an open-source architecture for neural machine translation. *arXiv preprint arXiv:2403.03582*.

Teresa Lynn. 2022. Report on the Irish language. https://european-language-equality.eu/deliverables/. Technical Report D1.20, European Language Equality Project.

Meta AI. Llama. https://www.llama.com/llama2/.

Microsoft Research. Microsoft Bing Translator. https://www.bing.com/Translator/.

Ailbhe Ní Chasaide, Neasa Ní Chiarán, Elaine Uí Dhonnchadha, Teresa Lynn, and John Judge. Digital Plan for the Irish Language Speech and Language Technologies 2023-2027. Available at https://assets.gov.ie/241755/e82c256a-6f47-4ddb-8ce6-ff81df208bb1.pdf.

Sharon O'Brien. 2004. Machine Translatability and Post-Editing Effort: How do they relate. In *Proceedings of Translating and the Computer 26*, London, UK. Aslib.

OpenAI. 2024. ChatGPT (November 29 version) [Large language model]. https://chatgpt.com/.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging GPT-4 for Automatic Translation Post-Editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023"*, pages 12009–12024, Singapore. Association for Computational Linguistics.

Charlie Robinson-Jones and Ydwine Scarse. 2022. *Report on the West Frisian Language (Language Technology Support of Europe's Languages in 2020/2021 - European Language Equality project)*.

Sanjun Sun. 2015. Measuring translation difficulty: Theoretical and methodological considerations. *Across Languages and Cultures*, pages pp. 29–54.

Natasha Tokowicz. 2014. Translation ambiguity affects language processing, learning, and representation. In *Selected Proceedings of the 2012 Second Language Research Forum*, pages pp. 170–180.

Nancy Underwood and Bart Jongejan. 2001. Translatability checker: a tool to help decide whether to use MT. In *Proceedings of Machine Translation Summit VIII*, Santiago de Compostela, Spain.

Bram Vanroy, Orphée De Clerq, and Lieve Macken. 2019. Correlating process and product data to get an insight into translation difficulty. *ERSPECTIVES-STUDIES IN TRANSLATION THEORY AND PRACTICE*, pages pp. 924–941.

Abigail Walsh, Teresa Lynn, and Jennifer Foster. 2019. Ilfhocail: A lexicon of Irish MWEs. *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages pp. 162–168.

Yangjian Wu and Gang Hu. 2023. Exploring Prompt Engineering with GPT Language Models for Document-Level Machine Translation: Insights and Findings. In *Proceedings of the Eighth Conference on Machine Translation*, pages 166–169, Singapore. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.