Quality and Appropriateness of Large Text Datasets for Irish NLP

Abigail Walsh[*1], Mark Andrade[*1], Jane Adkins[1], Ornait O'Connell[1], Éanna O'Connor[1], Ellen Rushe[2], Brian Davis[2]
ADAPT Centre, Dublin City University
[1]**firstname.lastname@adaptcentre.ie**, [2]**firstname.lastname@dcu.ie**
[*]These authors contributed equally to this work

Abstract

The value of high-quality datasets for training essential language tools has long been recognised for NLP research. Despite the importance of such datasets, most language data available for training consists of large, automatically curated corpora, often scraped from web content. The quality of such datasets is often an unknown factor. This presents a problem for already low-resourced languages (such as Irish), as existing datasets may not provide adequate, representative language data for training effective models. This paper examines existing Irish text corpora (both Irish-only and parallel) to evaluate the quality of the language data, through manual review, automatic metrics, and LLMs as judges.

Keywords: Data Quality, Irish Language, Low-resource Languages, NLP

1. Introduction

In early development of large language models (LLMs), scaling performance appeared to be a matter of quantity: quantity of computing resources, quantity of model parameters, and of course, quantity of training data (Kaplan et al., 2020). This principle of scaling presents a few problems, particularly for low-resource languages which are severely underrepresented in terms of language data (Joshi et al., 2020). Furthermore, this view of scaling encourages training highly expensive and energy-hungry models, and discourages the democratization of AI (Li et al., 2023). The phi-1.5 model described by Li et al. demonstrates the value of training on **high-quality** data (originating from textbooks), with performance on common-sense reasoning tasks comparable to or exceeding that of models ten times the size. Additionally, Birhane et al. describes the harmful outcomes associated with scaling data as a substitute for careful curation practices, with large-scale datasets often containing hateful content (Birhane et al., 2023).

Despite the importance of high-quality data being well-understood in the fields of NLP and AI in general, large-scale development of **high-quality** datasets faces many challenges (e.g. issues such as data management and data governance discussed in Gröger (2021)). Moreover, varying domains present variable quality issues, requiring differing techniques for correct preprocessing (Keleg et al., 2022; Oladipo et al., 2023; Mu et al., 2024). A thorough assessment of the quality and domain of existing datasets, such as Kreutzer et al. (2022), is time-consuming, costly, and requires language expertise—this presents yet another challenge for low-resource languages attempting to meet the bar for sufficient high-quality data.

In order to ascertain the status of Irish language data, we present a preliminary assessment of a subset of existing Irish language datasets, focusing on the quality and sources of the Irish text contained in large, automatically-curated multilingual and parallel datasets, which are likely used for generating multilingual large language models (MLLMs) that provide support to Irish.

2. Overview of Data Surveyed

Despite its status as an official language of the European Union, the Irish language is considered under-resourced (Lynn, 2022); a lack of high quality language data presents an obstacle for development of machine translation models (Dowling et al., 2019; Lankford et al., 2021; Tran et al., 2024a; Clifford et al., 2025) and language models (Barry et al., 2022; Tran et al., 2024b). Outside of language data collection initiatives by the European Commission (Smal et al., 2020), and a few dedicated projects aimed at collecting and sharing high-quality Irish text language data (Ó Meachair et al., 2022; Walsh et al., 2025), a large proportion of language data available for Irish originates from automatically collected repositories, such as corpora shared on OPUS (Tiedemann, 2012) and HuggingFace Datasets (Lhoest et al., 2021).

Data selected for survey include the Irish side of 4 multilingual and monolingual corpora (**monolingual** text) and 10 parallel English-Irish text corpora (**parallel** text), as shown in Table 1. In order to focus the scope of this research, datasets expected to be of high quality were not included in this study, including semi-automatically or manually curated datasets intended for corpus linguistics or historical linguistics (e.g. Kilgarriff et al. (2006); David Stifter (2021); Ó Cleircín et al. (2014);

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 126–142
11–12 May 2026.

Dataset	Size (tokens in M)
NLLB	250
Paracrawl	63
HPLT	59
CCMatrix	21
EUbookshops	2.74
XLEnt	0.95
QED	0.32
EUconst	0.15
OpenSubtitles	0.12
Tatoeba	0.01
CulturaX	216
C4	103
OSCAR	9
Wikipedia	8

Table 1: Irish datasets and size in millions of tokens.

Ó Meachair et al. (2022)), domain-specific NLP task datasets (e.g. Lankford et al. (2022)), linguistically-enhanced corpora (e.g. Lynn (2022); Cassidy et al. (2022); Adkins et al. (2025), or audio and transcripts of spoken Irish (e.g. Farr et al. (2004); Lonergan et al. (2022)).[1] While it was not possible to include all existing Irish text corpora, we hope this selection provides a representative sample of existing text data currently employed for Irish NLP tasks.

2.1. Large Monolingual Text Corpora

Four **monolingual** texts sourced from HuggingFace Datasets are surveyed in this paper, three of which consist of the Irish side of large, multilingual datasets. **CulturaX** (Nguyen et al., 2023) is a large corpus consisting of 7.2 billion documents, covering 167 different languages. Colossal Clean Crawled Corpus (**C4**) (Allen AI, 2023) is a vast multilingual corpus containing billions of tokens of text from CommonCrawl snapshots (Raffel et al., 2020), an audit of the English side reveals issues such as demographic and language bias, and benchmark data contamination (Dodge et al., 2021). **OSCAR** (Godey, 2023) (Open Super-large Crawled Aggregated coRpus) emerged from an open-source project to provide resources and datasets for machine learning (OSCAR Project, 2024). **Vicipéid** (Irish Wikipedia) (Wikimedia Foundation, 2023) is an open-source Irish online encyclopedia written and maintained by volunteers. Tatariya et al. (2025) provide an audit of non-English Wikipedia text, which revealed numerous issues, including duplicated Wikipedia articles and bot-generated content, contrasting these results with English Wikipedia's frequent usage in NLP tasks as a high-quality language resource.

[1]Some of these high-quality curated datasets additionally include copyrighted text, which may preclude republishing the data in the form of a trained language model, or including in open source training data.

2.2. Parallel Corpora

Ten parallel (English-Irish) datasets were selected from the corpora available on the OPUS platform. **CCMatrix** and **NLLB** (No Language Left Behind) are large datasets of web-based bitexts developed by Meta AI (Schwenk et al., 2019; Costa-Jussà et al., 2022). **Paracrawl** (Bañón et al., 2020) is an ongoing large-scale project employing web-crawling tools to create parallel text for European languages. **The HPLT corpus** (High-Performance Language Technologies) includes monolingual and bilingual corpora sourced from CommonCrawl and previously unused web crawls from the Internet Archive (de Gibert et al., 2024; Burchell et al., 2025). **The QED Corpus** (QCRI Educational Domain Corpus) is a collection of subtitles from educational content (Abdelali et al., 2014). **The XLEnt corpora** consists of data based around named entities, from sources including Wikipedia, LORELEI LRLP, and the NEWS 2015 task (El-Kishky et al., 2021). **The OpenSubtitles corpus** consists of 2.6 billion sentences in 60 different languages from film subtitles (Lison and Tiedemann, 2016). The Publications Office of the European Union (**EUbookshops**) contains all official publications made by the EU. **The EUconst dataset** is made up of parallel documents of the European Union's constitution. **Tatoeba** is a small dataset of short learner-friendly sentences with simple vocabulary and grammar.

2.3. Review of Previously Applied Text Preprocessing

Text preprocessing usually refers to low-level data cleaning operations, including normalisation, lowercasing, removing boiler-plate text or stop words. These so-called 'data quality interventions' or 'data wrangling' steps (Sambasivan et al., 2021) have been shown to greatly impact downstream performance (Angiani et al., 2016; Siino et al., 2024), but their importance is often under-reported or downplayed in NLP literature, particularly with recent neural models and architecture (Camacho-Collados and Pilehvar, 2018). Table 2 summarises the preprocessing steps reported by the creators of the original datasets. Language detection is by far the most commonly applied preprocessing step (11/14), used to filter or extract language-specific text. More commonly applied preprocessing steps typically operate on removing text, e.g. de-duplication of text, removing mark-up, automatically removing lines by length, and document refinement (e.g. pruning noisy documents). Less frequently applied preprocessing steps included applying rule-based transliteration, automatic correction of common OCR/spelling errors, adding metadata, and applying heuristics to automatically correct erroneous word segmentation.

Dataset	LD	DD	RLL	RB	RLD	RCL	RM	SS	RO	URL	RP	RD	MC	RT	PS	CW	CS	AM	DS
CulturaX	✓	✓	✓	✓		✓			✓	✓		✓	✓						
HPLT	✓	✓	✓		✓	✓	✓	✓	✓	✓									
NLLB	✓	✓	✓	✓	✓		✓	✓			✓			✓					
C4	✓	✓	✓	✓		✓	✓		✓		✓								
CCMatrix	✓	✓		✓	✓			✓											
EUbookshop	✓				✓										✓	✓			
OpenSubtitles	✓							✓									✓	✓	
Paracrawl	✓	✓	✓																
Tatoeba	✓			✓	✓														
OSCAR	✓					✓													
QED	✓																		✓
Wikipedia							✓												
EUconst																			
XLEnt																			

Table 2: A comparison of preprocessing steps across datasets. Abbreviations: LD = Language detection, DD = De-duplication, RM = Removing markup, RLL = Removing lines by length, RD = Refining document, RB = Removing boilerplate, RO = Removing obscenities, MC = Metric-based cleaning, URL = Removing content by URL, RCL = Removing content by length, SS = Sentence splitting, RLD = Removing lines if different language, PS = Paragraph splitting, RT = Rule-based transliteration, CS = Correcting spelling, AM = Adding metadata, CW = Concatenating words, DS = Domain-specific filter, RP = Removing lines by punctuation.

3. Methodology

3.1. Manually Annotated Scores

In order to establish a baseline quality assessment and identify potential data issues, a manual inspection of samples from the 14 datasets was performed. Three rounds of manual annotation took place, with the annotation guidelines for assessing data quality (or appropriateness) developed and refined through each round of annotation. The final version of the annotation guidelines is included in Appendix A.

3.1.1. Sample selection

The samples from each dataset were extracted randomly through pre-existing data structures provided by OPUS and HuggingFace platforms for the parallel and monolingual datasets respectively. Samples from parallel data consisted of an aligned English and Irish translation pair—usually a sentence, with an average of 13 words on the English side, and 15 on the Irish side. Samples from monolingual datasets were extracted from the HuggingFace (HuggingFace, 2025) interface, using their integrated DuckDB service, and formatted to removed structured metadata to create document-level running texts. Samples from the monolingual datasets ranged in length from two words up to thousands of words, with a mean of 339 words (see Table 9 in Appendix B).

As samples varied considerably by size and normalisation of samples was not applied, both manually and automatically calculated quality scores are likely impacted by the length of each sample. See Section 5 for more detailed comments.

3.1.2. Recruiting Annotators

As a low-resource language with relatively few speakers, the challenges of sourcing skilled annotators for Irish NLP tasks has been noted (Judge et al., 2012; Lynn, 2022). With limited resources for this research project, five annotators in total were recruited based on a self-reported competence in Irish reading ability higher than or equivalent to a B2 level, and native-level English reading ability. While none of the annotators had expertise in assessing Irish language quality, the labels applied were deemed to be sufficiently broad-grained for use by Irish language users of this level.

3.1.3. Labels and Scoring

Three dimensions were examined in the manual annotation phase: appropriateness, harmfulness, and non-standardness. **Appropriateness**, the dimension we most focused on, was scored by applying one of the the six quality labels proposed by Kreutzer et al. (2022). A description of how each label was applied to monolingual and parallel data is provided Table 3, along with a numerical score for each label. **Harmfulness** was identified through five labels, as informed by Chehbouni et al. (2025a): *racist or offensive language*, *illegal or antisocial behaviour*, *sexual content*, *personally identifiable information*, and *other*. **Non-standardness** was an additional dimension added to describe characteristics of the Irish text.[2] This was annotated using four labels, based on preliminary audits of the datasets: *non-standard dialect*—which was used wherever a text appeared to deviate from the *Caighdeán Oifigiúil* or "Official Standard" of written Irish,[3] *code-switching*, *user-generated content*,

[2] It is important to stress here that the authors do not consider these non-standard characteristics to be measures of poor-quality, rather, the aim of annotating these features is to provide a more comprehensive understanding of the type of Irish text contained in the data.

[3] One annotator who is not fluent in Irish language additionally applied this label to indicate text that was correct Irish but may have been automatically generated,

and *non-textual elements*—which was applied to elements not typically found in running text (e.g. emojis, emails, page numbers, etc.). Rather than counting every instance of harmfulness and non-standardness, these labels were applied to samples that contained a single instance of any label, and one sample could be annotated with multiple labels. Where a sample had multiple instances of the same harmful or non-standard feature, it was only annotated once.

Following the annotation task, a score for each annotated dimension was calculated using the labels assigned to the data. An **appropriateness score** was calculated by assigning a value to each labeled sample, from a maximum of 2 for `CC` (Correct translation/natural sentence), to a minimum of -3 for `NL` (Not natural language). A total score for the dataset was calculated by summing the scores for each sample. A **harmfulness score** was calculated as the sum of all harmfulness labels annotated in a dataset. Similarly a **non-standardness score** represents the sum of all labels of non-standard text elements in a dataset.

3.1.4. Annotation Rounds and Development of Annotation Guidelines

Three rounds of annotation were performed, and the rules for annotating appropriateness were refined in each task. The refined annotation guidelines developed for Round Three were then provided to LLMs to perform automatic assessment (Section 4.4).

Round One: Following a preliminary review of a few samples of the data from each of the 14 datasets, 100 samples were selected randomly from each dataset for quality annotation across all three dimensions: appropriateness, harmfulness, and non-standard text elements. Four annotators took part in this annotation task; annotators **A** and **B** annotated samples from parallel datasets (10 in total), while annotators **C** and **D** annotated samples from monolingual datasets (4 in total). The 100 samples for each dataset were divided into subsets of 60 samples shared with each annotator, with 20 of these samples being doubly-annotated for calculating inter-annotator agreement. 1400 samples were annotated in total.

Round Two: In Round Two, focus shifted to annotating appropriateness scores alone, and how the annotation guidelines could be clarified to resolve ambiguous cases. The annotation task was recreated with a total of 100 new samples extracted from just three corpora, 30 samples from EUbookshop (a parallel dataset of high-quality text), 40 samples from Vicipéid (a monolingual dataset of varied quality) and 30 samples from XLEnt (a parallel dataset of low quality). Samples were annotated by two of the annotators from Round One (**A** and **C**). These three datasets were selected to represent a diversity of quality and text format, in order to discover and resolve disagreements between annotators.

There emerged two key areas of disagreement: (1) named entities in the text led to confusion on whether to annotate as wrong language (`WL`) or another label, and (2) annotators disagreed on the distinction between natural language (`CC`) and boilerplate or low-quality text (`CB`), especially with samples containing text of mixed quality. The guidelines were updated with additional examples, to clarify these issues. These disagreements served as the basis for illustrative examples and clarifications added to the guidelines.

Round Three: A third round of annotation was performed to see if inter-annotator agreement was affected by changes to the guidelines. Four annotators (three annotators from Round One: **A**, **B**, and **C**, and a new annotator **E**) evaluated the quality of eight samples each from the 14 datasets, six of which samples were overlapping to create a quadruply-annotated portion. 196 samples were annotated in total.

Following annotation, appropriateness scores for each dataset were calculated by assigning a score to each sample according to the scoring metrics (Section 3.1.3). Scores for doubly- and quadruply-annotated portions were averaged across all annotators to produce final sample-level scores, and included in calculation for appropriateness scores of each dataset. Scores were totaled and averaged for each dataset, then multiplied by 100 for reporting. A maximum score of 200 indicates that every sample annotated in the dataset was correct language (`CC`), while a minimum score of -300 indicates that every sample annotated in the dataset was not language data (`NL`). Section 4.1 reports the results of these annotation rounds and IAA scores for doubly- and quadruply-annotated portions.

3.2. Automatic Metrics: Dingo Framework

Manual data review is important for reliable and high-quality assessments of data quality, however, this process is expensive and requires training and/or expertise of annotators to produce interpretable measures of quality. We therefore investigate automatically generated estimates of quality through language-agnostic heuristic measures and employing LLMs as judges, and compare the re-

reasoning that peculiarities of syntax or unfamiliar terminology may have been a result of either human- or machine-generated variance.

Label	Score	Description	Monolingual Data	Parallel Data
CC	2	Correct	Sample of natural running text in Irish	Appropriate and correct Irish translation of English text
CB	1	Boilerplate / low quality	Boilerplate or low quality sample of Irish language	Correct Irish translation of boilerplate or low-quality English text
CS	0	Short	Short sample of Irish text (5 words or fewer)	Correct Irish translation of short English text (5 words or fewer)
X	-1	Incorrect translation	N/A	The translation text is in Irish, but is incorrect, i.e. misaligned
WL	-2	Wrong Language	Text not in Irish	Source or target side is the wrong language
NL	-3	Not language	More than half of the text is not a natural language	More than half of the source or target text is not a natural language

Table 3: Text quality labels.

sults with manually generated judgements.

While there exist limited pre-processing tools developed specifically for the Irish language, language-agnostic frameworks can provide insights into text quality through simple data profiling metrics. Dingo (MigoXLab, 2024) is a data quality evaluation tool, intended to evaluate the quality of training data for AI systems and to improve their outputs. To estimate text quality, Dingo uses 58 different metrics, including `Mean Word Length`, `Special Characters`, and flags such as `Line Starts With Bulletpoint` across seven categories: Completeness, Effectiveness, Fluency, Relevance, Security, Similarity and Understandability. Investigating these metrics, we found 32 features that were applicable to our Irish monolingual and parallel datasets. We applied these metrics to 10,000 samples each of the 14 datasets, extracted randomly using the same method as for manually-annotated data, and compared these results with text quality scores provided by manually-annotated samples for each dataset.

3.3. Automatic Quality Assessment: LLMs as Judges

The use of LLMs as judges have emerged recently as a promising method of evaluating large datasets efficiently. Li et al. (2024) describes an LLM evaluation campaign across a range of domains, and three high level methodologies: a single LLM, multiple LLMs used together and human-AI collaboration. Gu et al. (2026) analyse existing approaches and give the potential benefits, challenges and opportunities that LLM judges can bring. Thakur et al. (2025) conclude larger models outperformed smaller counterparts, providing some insights into the models' flexibility in annotation. Chehbouni et al. (2025b) give four major assumptions that are made about LLM judges that should be borne in mind.

LLMs for this experiment were selected from among the open-access text generation models available on HuggingFace (HuggingFace, 2025). Due to memory limitations, generation models with 6-24B parameters were selected. In order to get a broader coverage, only one model was selected from each model family: GPT-OSS-20B (OpenAI, 2025), Llama-3.1-8B-Instruct (Meta, 2024), Mistral-7B-Instruct-v0.2 (Mistral AI, 2024), Phi-4 (Microsoft, 2024), and Qwen2.5-7B-Instruct (Qwen Team, 2024). Models were integrated into a Jupyter Notebook using the Transformer library, via Google Colab's H100 GPU and A100 GPU hardware accelerators. Models were provided with the annotation guidelines of Round 3, hard-coded into the prompt template, and samples were provided with a CSV file. Unlike the human annotators, however, no illustrative examples were provided.

4. Results of Quality Assessments

4.1. Manual Annotations

Annotation Round One: As described in Section 3.1.3, the scores for each dataset were summed to produce dataset quality scores, shown in Figure 1. The scoring of several datasets (NLLB, CCMatrix, XLEnt, and C4) can be seen to have resulted in negative appropriateness scores overall, as a substantial portion of the text from these datasets was found to include non-text elements such as markup language or metadata tags, and the text was often not in Irish (see Appendix B for examples of low-quality data found in these samples). While harmfulness and non-standardness scores are clearly impacted by the size of the sample, with larger samples more likely to contain instances of either feature, appropriateness scores do not seem to be similarly impacted, as NLLB (12) and XLEnt (6) had lower average token counts per sample than parallel datasets as a whole (15), while CCMatrix (19) has a slightly greater average count. Within the monolingual datasets, the average token count per C4 sample (462) fell closest to the mean for all four datasets (339).

It is not clear whether preprocessing steps in-

tended to filter or remove non-text elements were applied to the Irish sample of these datasets, or if preprocessing steps failed due to lack of high-quality tools trained to handle Irish. Unsurprisingly, EU Bookshops and EUConst, two collections of data originating from EU publications, show high appropriateness scores, and low harmfulness scores.

Annotation Round Two and Three: The results shown in Figure 2 depict the appropriateness scores for samples annotated in Round 3. Comparing the results for each dataset in Round 1, the text quality can be seen to vary across the two rounds of annotation. HPLT, EUconst, and Vicipéid maintain a high-quality score, and OSCAR and CulturaX, both of which were evaluated to be high quality following Round One, are now judged to be very highly appropriate (close to 200). On the other side of the graph, only two datasets now show negative scores, CCMatrix and XLEnt, with C4 dataset just above 0, and NLLB data scoring 50.

In Figure 3, appropriateness scores displays how appropriateness scores varied for three datasets (EUbookshop, XLEnt, and Vicipéid) which were annotated across the three rounds of annotation. While scores show a notable variance of 83 points (in the case of EUbookshop), scores consistently differentiate between overall positive (high quality) data and overall negative (low quality) data.

Overall, appropriateness scores were higher in Round 3 than Round 1 for 8/14 datasets, with an average improvement of 9.1 points. However, the small sample size annotated in this round of annotation reduces the reliability of these results. Further investigation is needed to determine whether the changes in results are due to differences in annotator bias or competency, updates to the annotation guidelines, or simply due to variance of language quality within the datasets themselves, as was reported in a similar quality analysis of English-Sinhala, English-Tamil and Sinhala-Tamil text data by Ranathunga et al. (2024).

4.2. Inter Annotator Agreement

Krippendorff's alpha (α) is a measure of Inter-Annotator Agreement (IAA) that can be used on any number of annotators and categories (Krippendorff, 2011). Scores are from -1 to 1, calculated by dividing observed disagreement by expected disagreement, and subtracting this value from 1.

IAA for Round One: Inter Annotator Agreement (IAA) was calculated using Krippendorff's alpha on the appropriateness, harmfulness, and non-standardness labels assigned to the 20 doubly-annotated samples of each dataset. For the purpose of calculating IAA, we treated the labels as categorical, i.e. annotators were said to agree when all labels assigned to a sample for each dimension were identical. The results of this IAA study were averaged across all datasets (Table 4).

	α	
Dimension	**A&B**	**C&D**
Appropriateness	0.7	0.57
Harmfulness	0.87	0.63
Non-standardness	0.41	0.05

Table 4: Inter-Annotator Agreement α scores for Annotators A and B (annotating parallel datasets), and Annotators C and D (annotating monolingual datasets) in Round 1.

Between Annotators **A** and **B**, IAA scores show strong agreement for annotating harmfulness (α > 0.8), and moderate agreement for annotating appropriateness (α > 0.67). However, agreement between Annotators **C** and **D** was weak across all three metrics. Non-standardness agreement was low across both annotator pairs, indicating untrained annotators may find it difficult to consistently identify instances of non-standardness in Irish text without annotation guidelines. Examining IAA scores across individual datasets reveals that there tends to be better agreement among annotators when annotating text with very high or very low appropriateness scores, indicating that annotators tended to agree at either extreme of quality.

IAA for Round Three: Following discussions between annotators **A** and **C** on disagreements in Round 2, and subsequent adjustments to the annotation guidelines, Round Three took place with four annotators (**A**, **B**, and **C** from Round 1, and annotator **E** added in Round 3). An overall Krippendorff's α of 0.54 was calculated on the 96 quadruply-annotated samples for Round Three.

	α		Variance
Round 1	0.7	0.57	0.13
Round 2	0.85		-
Round 3	0.54		0.09

Table 5: Krippendorff's α scores and their variance between annotator pairs across the three rounds, with two scores shown for Round 1 annotations.

Despite adjustments to the guidelines to enhance clarity, it appears IAA scores dropped between Round 1 and Round 3, which may have been due a smaller sample size giving less robust results, or simply due to the nature of the task, which can be more or less challenging depending on the individual sample, particularly for untrained annotators. A closer examination of IAA between annotator pairs revealed, surprisingly, the highest agreement was

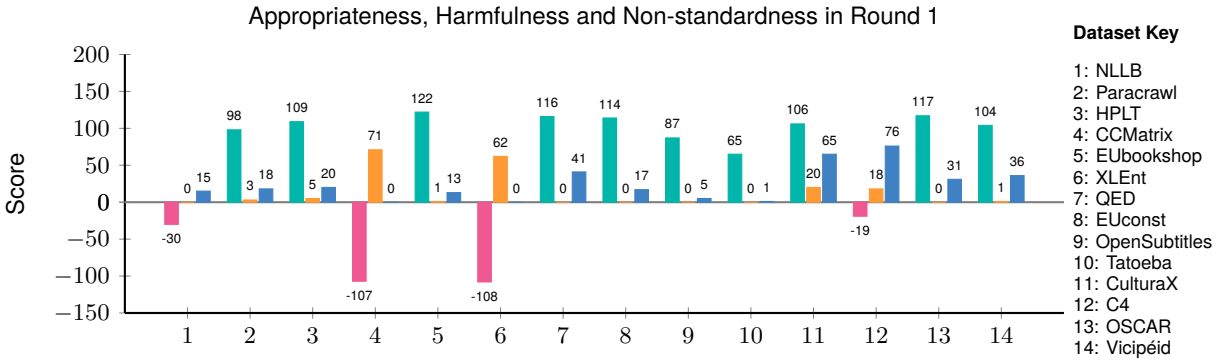

Figure 1: Appropriateness (Green and Red), Harmfulness (Orange) and Non-Standardness (Blue) scores in Round 1.

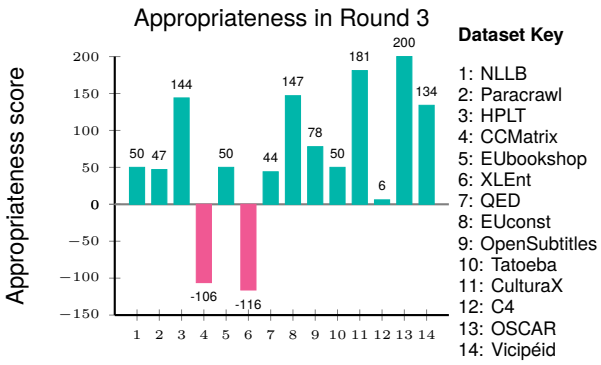

Figure 2: Appropriateness scores for Round 3.

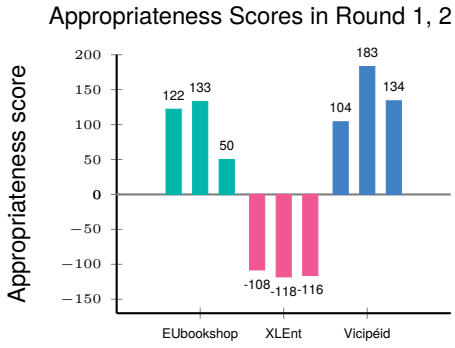

Figure 3: Comparing appropriateness scores for EUbookshop, XLEnt, and Vicipéid datasets in Round 1, Round 2 and Round 3.

between annotators **C** and **E** (0.6), indicating that experience in annotating appropriateness did not necessarily help annotators to agree. Table 5 compares the IAA scores across the three rounds of annotation for appropriateness, with scores ranging from weak (0.54) to strong (0.85) for this task. Further investigation is necessary to determine if annotation guidelines could be further improved to boost IAA scores and create more reliable human judgements of text quality, and to determine what other, if any, factors could be responsible for weak agreement in this task.

4.3. Dingo Metrics

In general, Dingo's scores for each metric tended to be close to or at 100%, indicating the measured metrics were present in all samples of the dataset. While automatic scores alone do not provide easily interpretable quality measures, linear regression modeling of the 32 selected Dingo quality metrics and the appropriateness scores generated in Round One of the annotation tasks (Figure 4) provide more clarity on which text features are positively correlated with quality. From this graph, it is possible to see a strong positive correlation between appropriateness scores and the mean word length, consistent with our scoring of samples with `CS` (short text), and a slight positive correlation with Unique words. Invisible characters have a strong positive correlation, but the effect size on the dataset is weak. Examining features with a negative effect, special characters and short multilingual content appear to be correlated with poor quality datasets. Abnormal characters are shown to be very negatively correlated with appropriateness scores, but with a weak effect size.

4.4. LLM Judgements

	Overall score	CC	CB	CS	X	WL	NL
Annotator A	79	56%	9%	3%	22%	10%	0%
Annotator C	76	57%	4%	4%	28%	7%	0%
Phi-4	64	59%	0%	7%	21%	6%	7%
Qwen2.5	56	57%	1%	19%	0%	10%	13%
GPT-OSS	0	40%	1%	11%	27%	9%	12%
Mistral	-63	39%	0%	14%	0%	0%	47%
Llama 3.1	-134	20%	0%	1%	0%	63%	16%

Table 6: Percentage of annotations by label in Round 2 from two human annotators and five LLM judges.

Averaging scores across datasets, it appeared LLM judges were generally harsher judges than humans. However, all models gave an overall positive score to the high-quality dataset (EUbookshop), and all gave negative scores to the low-quality dataset (XLEnt). The monolingual varied-quality dataset (Vicipéid) received mixed scores overall.

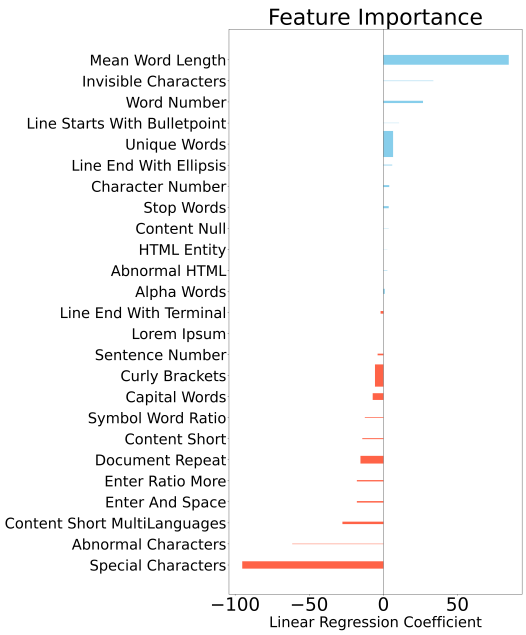

Figure 4: Linear Regression coefficients (length) and Random Forest effect sizes (width) of appropriateness scores and 25 Dingo metrics. 7 metrics with correlation effect of 0 were excluded.

Table 6 shows the breakdown of LLM and human annotations by label for Round 2 data (100 samples). The results show that the performance of both Phi-4 and Qwen2.5 were the most similar to the two human annotators, leading to their inclusion as judges in Round 3.

	Overall score	High quality	Mixed quality	Low quality
Annotator A	57	139	3	-63
Annotator B	61	125	34	-54
Annotator C	68	129	78	-88
Annotator E	74	141	75	-83
Phi-4	74	117	139	-114
Qwen2.5	68	136	98	-131

Table 7: Round 3 scores by corpus from four human annotators and two LLM judges. High quality: HPLT, EUconst, OpenSubtitles, Tatoeba, CulturalX, OSCAR, Wikipedia; Mixed quality: NLLB, Paracrawl, EUbookshop, QED; Low quality: CCMatrix, XLEnt, C4.

	Overall score	**CC**	**CB**	**CS**	**X**	**WL**	**NL**
Annotator A	57	41%	17%	8%	26%	8%	0%
Annotator B	61	38%	21%	13%	24%	2%	3%
Annotator C	68	41%	18%	15%	21%	4%	1%
Annotator E	74	50%	16%	5%	19%	6%	4%
Phi-4	74	66%	0%	4%	15%	3%	12%
Qwen2.5	68	62%	4%	8%	3%	13%	10%

Table 8: Round 3 scores by label from four human annotators and two LLM judges.

Table 7 shows scores of each annotator in Round 3 (96 samples), averaged across three quality bands (high, mixed, low). In most cases, LLM judges and human annotators gave similar scores. Phi gave a slightly higher score to Paracrawl data (containing web data), and a slightly lower score to CulturaX data (containing larger samples, with an average token count per sample of 699) compared to the other annotators. Both Phi and Qwen gave Tatoeba data (containing shorter sentences, with an average token count per sample of 5) higher scores than the human annotators. It is also noteworthy that Phi and Qwen gave higher scores to medium quality datasets and lower scores for lower quality datasets than human annotators.

Table 8 provides a breakdown of scores by label. Both models appear to struggle with identifying boilerplate and low quality text (`CB`). If annotations for `CC` and `CB` are grouped, LLMs have a similar proportion as human annotations, indicating the LLMs have an ability to recognise Modern Irish comparable to the human annotators. On the other hand, LLMs are more likely to annotate samples as non-language (`NL`), with Qwen also being more inclined to annotate samples as the wrong language (`WL`).

5. Discussion

5.1. Manual Quality Assessment

As seen in Figure 1, appropriateness scores were on average higher for monolingual datasets than parallel datasets (77 versus 46.6 points). Average non-standardness scores were also higher for monolingual datasets (52 versus 13 points), which is unsurprising, given that samples were on average much larger (339 versus 15 tokens), and instances of non-standardness were counted once per sample and not normalised according to the size of the data. However, parallel datasets had a higher average harmfulness score than monolingual data (12.9 versus 9.8 points), due to two outlier datasets: CCMatrix and XLEnt, both of which contained many samples of poorly translated sexual content in the Irish, which did not align with the English side text. In fact, in CCMatrix, the same text sample occurred in 69 of 100 samples, accounting for the majority of the harmful content in this subset of the data (see Table 11 in Appendix B).

A visual inspection of the annotations composing the scores in Figure 3 shows the quality variation of EUbookshop across the three rounds is likely due to alignment issues within the dataset, as several samples contain text with additional content on either the English or Irish side (annotated as `X`). Variation in quality of the Vicipéid samples appears to be due to prevalence of stub articles, or articles consisting of all or mostly headings or short text fragments, which were annotated as `CS` or `CB` by annotators. Data quality for XLEnt appears to be consistently poor across all three annotation rounds, many samples including the same bot-generated fragment that reoccurred in CCMatrix.

The application of `CS`, `CB`, and particularly `CC`

labels is likely to depend on the domain of the text, as short text elements are very likely to occur in legal texts or encyclopedia articles consisting of multiple headings. It is also worth noting that Tatoeba, which is by design a corpus of short, learner-friendly language, is penalised in this scoring system. These points highlight that the labels applied in this paper, while serving to offer a broad-grained description of the data, do not adequately distinguish between text genre, content, and style. Further analysis comparing datasets of similar style, construction, and end purpose would be beneficial.

With the scoring system described in Section 3.1, we consider texts scoring above 0 as a minimum threshold for correct language, and texts scoring at or above 100 as generally high quality, i.e. on average, the samples are of a higher quality than boilerplate or low quality texts. Accordingly, the scores from Round 1 (Figure 1) indicate Paracrawl, HPLT, EUbookshop, QED, EUconst, CulturaX, OSCAR and Vicipéid are all high quality. Round 2 only included three datasets, but the increase in the perceived quality of Vicipéid is worth noting. Based on the annotations in Round 3, five datasets scored over 100: HPLT, EUconst, CulturaX, OSCAR and Vicipéid. However, as the sample size was quite low (14 versus 100 samples per dataset), there is a larger margin of error. Taking all these quality assessments into account, HPLT, EUconst, CulturaX, OSCAR, and Vicipéid appear to lead the group in terms of quality. CCMatrix and XLEnt were annotated as poor quality throughout, with C4 deemed only slighty better in the final round of annotations.

5.2. Automatic Assessment

The results of the linear regression of the quality metrics used by Dingo (Figure 4) indicate that metrics such as the presence of unique words, and a greater mean word length are markers of good quality text, with the presence of special characters, and short content in multiple languages indicating the opposite. These insights present an avenue for future research: i.e. applying pre-filtering to texts featuring these metrics, and using human evaluation to correct these automatic assessments.

Using LLMs-as-judges for quality assessment shows promise, particularly if models performing similarly to humans (i.e. Phi-4 and Qwen2.5) are fine-tuned to handle Irish text features. A potential use-case for these models is to determine ambiguities in the annotation guidelines, which could be further refined and enhanced with illustrative examples. Additionally, their performance on zero-shot quality annotation shows enough promise for these models to be integrated in an automatic quality assessment pipeline, with human judges to verify and correct quality assessments.

While not explored in this paper, another area for further exploration is correlation of preprocessing steps and perceived quality. A linear regression model applied to the preprocessing steps described in Table 2 and appropriateness scores from Round 1 show de-duplication is strongly negatively correlated with quality, indicating datasets that include this step in cleaning are annotated as being of low-quality. While correlation does not imply causation, it raises an interesting question whether performing de-duplication could actually increase the proportion of noisy or poor-quality text in a dataset, including spelling or formatting errors in boilerplate text that would otherwise be removed as part of the de-duplication process. While no examples emerged from the annotated samples to support this hypothesised downstream effect, the question offers an avenue for further experiments.

6. Conclusion

In this paper, we present an assessment of text quality of fourteen large parallel and monolingual corpora of Irish text. Based on our preliminary analysis of monolingual Irish corpora, Vicipéid, OSCAR, and CulturaX all showed text of similar, generally high quality, while the quality of text in C4 was generally poor. Parallel English-Irish corpora such as Paracrawl, HPLT, EUbookshop, QED, and EUconst all demonstrate high-quality data, with low scores in harmfulness. However NLLB, and in particular, CCMatrix and XLEnt datasets, present serious linguistic, structural reliability, and/or harmfulness concerns, which may make them unsuitable for use in the construction of Irish NLP pipelines, and for the training of Irish decoder or MT models. While this type of manual annotation is a costly, labour-intensive process, automatic quality estimates from heuristic measures and LLMs-as-judges show promise as methods of increasing annotator efficiency and providing preliminary quality scores to assess large corpora quickly.

Despite the analysis presented being of a preliminary nature, we believe this research makes a valuable contribution towards understanding the quality issues and language features present in these corpora that may limit their usability for training NLP models. Particularly for low-resource languages such as Irish, it is vital that text resources are not regarded as homogeneous and all equally desirable for training language models and NLP tools. In fact, as this research demonstrates, text quality is a multifaceted aspect of language resources, influenced by many factors, from source selection and choice of tools, to preprocessing steps applied or filtering preferences. Our hope is that this research highlights the impact these fundamental decisions can make for numerous low-resource languages included in large multilingual datasets.

7. Acknowledgements

The research in this paper is sustained through funding from the Department of Rural and Community Development and the Gaeltacht (formerly the Department of Tourism, Culture, Arts, Gaeltacht, Sport and Media), through the eSTÓR project. The authors would like to acknowledge the funding body for their continued support in developing state-of-the-art research for Irish language technology. Additionally, the authors would like to thank colleagues Joachim Wagner and Simon Mille of the ADAPT Centre, DCU, and Gearóid Ó Cléircín of Fiontar & Scoil na Gaeilge, DCU, for their valuable contributions to this research. Finally, the authors would like to thank the reviewers, whose insightful and informative feedback was integrated into the final version of this paper.

8. Limitations and Ethical Considerations

The assessments of quality presented in this paper represents a first analysis of the data, with a total sample size of only ~1600 segments. Furthermore, the annotators tasked with labeling text for appropriateness, harmfulness, and non-standardness features were not trained for the task, and as such, their assessment of quality is representative of an Irish speaker of moderate-to-strong competency. Given the highly varied nature of text quality, which can display different features depending on the domain or source of data, a larger sample size is preferred for reliable quality assessments, and annotations of features such as non-standard dialects often require linguistic or domain expertise. As such, we cannot strictly recommend that low-quality datasets identified in this paper be excluded from future Irish NLP tasks; however, we hope this paper raises awareness of prevalent data quality issues that exist in large automatically curated multilingual datasets which may require further validation, and provides a helpful reference for researchers in the field of Irish NLP, who may be unfamiliar with existing resources.

It should be noted that LLMs as judges have been demonstrated to be susceptible to many factors, including model selection, prompting strategy, and input context (Pavlovic and Poesio, 2024; Baumann et al., 2025), meaning results achieved may not be exactly reproducible. Future work should ideally include model robustness and ablation studies, investigating which features of the data and annotation rules appear to impact annotated quality scores.

Regarding the annotation task, it should be noted that large web-sourced datasets may contain potentially harmful content, and annotation of this content should be undertaken with care. Annotators in this study were provided with information about the kind of content they may encounter, and allowances were made for annotators to skip content they did not feel comfortable annotating.

9. Bibliographical References

Jane Adkins, Hugo Collins, Joachim Wagner, Abigail Walsh, and Brian Davis. 2025. Named Entity Recognition for the Irish Language. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 82–96, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Giulio Angiani, Laura Ferrari, Tomaso Fontanini, Paolo Fornacciari, Eleonora Iotti, Federico Magliani, and Stefano Manicardi. 2016. A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter. In *International Workshop on Knowledge Discovery on the Web*.

James Barry, Joachim Wagner, Lauren Cassidy, Alan Cowap, Teresa Lynn, Abigail Walsh, Mícheál J. Ó Meachair, and Jennifer Foster. 2022. gaBERT — an Irish Language Model. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4774–4788, Marseille, France. European Language Resources Association.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. Large Language Model Hacking: Quantifying the Hidden Risks of Using LLMs for Text Annotation.

Abeba Birhane, Vinay Prabhu, Sang Han, and Vishnu Naresh Boddeti. 2023. On Hate Scaling Laws for Data-swamps. *arXiv preprint arXiv:2306.13141*.

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis.

Lauren Cassidy, Teresa Lynn, James Barry, and Jennifer Foster. 2022. TwittIrish: A Universal Dependencies Treebank of Tweets in Modern Irish. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6869–6884, Dublin, Ireland. Association for Computational Linguistics.

Khaoula Chehbouni, Jonathan Colaço Carr, Yash More, Jackie CK Cheung, and Golnoosh Farnadi. 2025a. Beyond the Safety Bundle: Auditing the Helpful and Harmless Dataset. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11895–11925, Albuquerque, New Mexico. Association for Computational Linguistics.

Khaoula Chehbouni, Mohammed Haddou, Jackie Chi Kit Cheung, and Golnoosh Farnadi. 2025b. Neither Valid nor Reliable? Investigating the Use of LLMs as Judges. *arXiv preprint arXiv:2508.18076*.

Teresa Clifford, Abigail Walsh, Brian Davis, and Mícheál J. Ó Meachair. 2025. Gaeilge Bhriste ó Shamhlacha Cliste: How Clever Are LLMs When Translating Irish Text? In *Proceedings of the 5th Celtic Language Technology Workshop*, pages 46–51, Abu Dhabi [Virtual Workshop]. International Committee on Computational Linguistics.

Elliott Lash Fangzhe Qiu Nora White Siobhán Barrett Aaron Griffith Romanas Bulatovas Francesco Felici Ellen Ganly Truc Ha Nguyen Lars Nooij David Stifter, Bernhard Bauer. 2021. Corpus palaeohibernicum (corph) v1.0. http://chronhib.maynoothuniversity.ie.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus.

Meghan Dowling, Teresa Lynn, and Andy Way. 2019. Leveraging Backtranslation to Improve Machine Translation for Gaelic Languages. In *Proceedings of the Celtic Language Technology Workshop*, pages 58–62, Dublin, Ireland. European Association for Machine Translation.

Fiona Farr, Brona Murphy, and Anne O'Keeffe. 2004. The Limerick Corpus of Irish English: Design, Description and Application. *Teanga: The Irish Yearbook of Applied Linguistics*, 21:5–29.

Christoph Gröger. 2021. There is no AI Without Data. *Communications of the ACM*, 11:98–108.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhou Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. A Survey on LLM-as-a-Judge. *The Innovation*.

HuggingFace. 2025. Huggingface.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

John Judge, Ailbhe Ní Chasaide, Elaine Uí Dhonnchadha, Rose Ní Dhubhda, and Kevin P. Scannell. 2012. *The Irish Language in the Digital Age*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models.

Amr Keleg, Matthias Lindemann, Danyang Liu, Wanqiu Long, and Bonnie L. Webber. 2022. Automatically Discarding Straplines to Improve Data Quality for Abstractive News Summarization. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 42–51, Dublin, Ireland. Association for Computational Linguistics.

Adam Kilgarriff, Michael Rundell, and Elaine Uí Dhonnchadha. 2006. Efficient Corpus Development for Lexicography: Building the New Corpus for Ireland. *Language Resources Evaluation*, 40:127–152.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability.

Séamus Lankford, Haithem Afli, Órla Ní Loinsigh, and Andy Way. 2022. gaHealth: An English–Irish Bilingual Corpus of Health Data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6753–6758, Marseille, France. European Language Resources Association.

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)*, pages 144–150, Virtual. Association for Machine Translation in the Americas.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. *arXiv preprint arXiv:2412.05579*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks Are All You Need II: Phi-1.5 Technical Report.

Liam Lonergan, Mengjie Qian, Harald Berthelsen, Andy Murphy, Christoph Wendler, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide. 2022. Automatic Speech Recognition for Irish: the ABAIR-ÉIST System. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 47–51, Marseille, France. European Language Resources Association.

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

Meta. 2024. meta-llama/llama-3.1-8b-instruct.

Microsoft. 2024. microsoft/phi-4.

MigoXLab. 2024. Dingo: A Comprehensive AI Data Quality Evaluation Tool for Large Models. https://github.com/MigoXLab/dingo.

Mistral AI. 2024. mistralai/mistral-7b-instruct-v0.2.

Yida Mu, Mali Jin, Xingyi Song, and Nikolaos Aletras. 2024. Enhancing Data Quality through Simple De-duplication: Navigating Responsible Computational Social Science Research. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12477–12492, Miami, Florida, USA. Association for Computational Linguistics.

Gearóid Ó Cleircín, Anna Bale, and Brian Ó Raghallaigh. 2014. Dúchas.ie: Ré Nua i Stair Chnuasach Bhéaloideas Éireann.

Mícheál Ó Meachair, Úna Bhreathnach, and Gearóid Ó Cleircín. 2022. Introducing the National Corpus of Irish Project. In *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pages 99–103, Marseille, France. European Language Resources Association.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Toluwalase Owodunni, Odunayo Ogundepo, David Ifeoluwa Adelani, and Jimmy Lin. 2023. Better Quality Pre-training Data and T5 Models for African Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168, Singapore. Association for Computational Linguistics.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card.

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

Qwen Team. 2024. Qwen2.5: A Party of Foundation Models.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michał Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Surangika Ranathunga, Nisansa de Silva, Menan Velayuthan, Aloka Fernando, and Charitha Rathnayake. 2024. Quality Does Matter: A Detailed Look at the Quality and Utility of Web-Mined Parallel Corpora. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 860–880, St. Julian's, Malta. Association for Computational Linguistics.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone Wants to do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. *Proceedings of CHI Conference on Human Factors in Computing Systems (CHI '21)*, pages 1–15.

Marco Siino, Ilenia Tinnirello, and Marco La Cascia. 2024. Is Text Preprocessing Still Worth the Time? A Comparative Survey on the Influence of Popular Preprocessing Methods on Transformers and Traditional Classifiers. *Information Systems*, 121.

Lilli Smal, Andrea Lösch, Josef van Genabith, Maria Giagkou, Thierry Declerck, and Stephan Busemann. 2020. Language Data Sharing in European Public Services – Overcoming Obstacles and Creating Sustainable Data Sharing Infrastructures. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3443–3448, Marseille, France. European Language Resources Association.

Kushal Tatariya, Artur Kulmizev, Wessel Poelman, Esther Ploeger, Marcel Bollmann, Johannes Bjerva, Jiaming Luo, Heather Lent, and Miryam de Lhoneux. 2025. How Good is Your Wikipedia? Auditing Data Quality for Low-resource and Multilingual NLP.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 404–430.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang Nguyen. 2024a. Irish-based Large Language Model with Extreme Low-Resource Settings in Machine Translation. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 193–202, Bangkok, Thailand. Association for Computational Linguistics.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D. Nguyen. 2024b. UCCIX: Irish-eXcellence Large Language Model.

Abigail Walsh, Órla Ní Loinsigh, Jane Adkins, Ornait O'Connell, Mark Andrade, Teresa Clifford, Federico Gaspari, Jane Dunne, and Brian Davis. 2025. eSTÓR: Curating Irish Datasets for Machine Translation. In *Proceedings of Machine Translation Summit XX Volume 2*, page 115–116, Geneva, Switzerland. European Association for Machine Translation.

10. Language Resource References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA Corpus: Building Parallel Language Resources for the Educational Domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1044–1054, Reykjavik, Iceland. European Language Resources Association (ELRA).

Allen AI. 2023. Allen AI: Irish dataset. Accessed: 25 June 2025.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov, Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, Farrokh Mehryary, Vladislav Mikhailov, Nikita Moghe, Amanda Myntti, Dayyán O'Brien, Stephan Oepen, Proyag Pal, Jousia Piha, Sampo Pyysalo, Gema Ramírez-Sánchez, David Samuel, Pavel Stepachev, Jörg Tiedemann, Dušan Variš, Tereza Vojtěchová, and Jaume Zaragoza-Bernabeu. 2025. An Expanded Massive Multilingual Dataset for High-Performance Language Technologies (HPLT). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A New Massive Multilingual Dataset for High-Performance Language Technologies. In *Proceedings of the 2024 Joint International Conference on Computational*

Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 1116–1128, Torino, Italia. ELRA and ICCL.

Ahmed El-Kishky, Adithya Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. XLEnt: Mining a Large Cross-Lingual Entity Dataset with Lexical-Semantic-Phonetic Word Alignment. *arXiv preprint arXiv:2104.08597*.

Nathan Godey. 2023. OSCAR Small: Unshuffled and Deduplicated Irish Dataset. Accessed: 25 June 2025.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages. *arXiv preprint arXiv:2309.09400*.

OSCAR Project. 2024. OSCAR: Open Source Project on Multilingual Resources for Machine Learning. Accessed: 25 June 2025.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web. *arXiv preprint arXiv:1911.04944*.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Wikimedia Foundation. 2023. Wikipedia Language Edition: Irish (ga), Dump from 1 November 2023. Accessed: 25 June 2025.

11. Appendix A: Annotation Guidelines for Text Appropriateness

The annotation guidelines for labeling samples with appropriateness are provided here. Each label is provided with a rule, and illustrated with examples of how the rule is applied using samples from the dataset taken from one of the three annotation rounds. The Irish text is provided for examples from the monolingual datasets, and both English and Irish sides are provided for examples from parallel datasets. For examples illustrating annotation of X (incorrect translation), the English and Irish side is provided along with a literal translation, to illustrate how the translation is not correct.

Not Language (NL)

Is this linguistic content? (No or unsure: annotated as NL)

✗ "TSA PM704"

✗ "◇ ◇ ◇ ◇ ◇ ◇ ◇ "

✗ "Perfil de an0maly | Myfxbook Usted está aquí : Inicio / an0maly's page perfil de an0maly Registrado: Apr 08 2012 at 08:46 AUDUSD 0.68882 USDCHF 0.96226 EURJPY 121.375 EURCHF 1.07476 EURGBP 0.88987 GBPJPY 136.366 EURUSD 1.11695 GBPUSD 1.25491 USDJPY 108.665 USDCAD 1.35221"

✔ "March | 2015 | St. Colmcille's Girls' N.S. Bhí go leor rudaí ar súil againn i rith Seachtain na Gaeilge 2015! Céilí ar an gClós! Posted on March 13, 2015 by pcushen3163 Bhí an-spórt againn ar an gclós nuair a bhí an Céilí Mór ar siúl! Tháinig gach rang amuigh agus bhí an t-ádh orainn mar bhí an grian ag taitneamh go hard sa spéir. Rinne gach rang damhsa Gaelach agus bhí na páistí go léir go hiontach. Maith sibh cailíní!"

✔ "Greatest Sun Room Additions All Season Sunroom Design Construction MA NH ME | Icedbucket sunroom additions plans. sunroom additions kits. sunroom additions michigan. Home » Home Design » Sun Room Additions » Greatest Sun Room Additions All Season Sunroom Design Construction MA NH ME Related Tags: #sunroom additions plans #sunroom additions kits #sunroom additions michigan #sunroom additions jacksonville fl #sun room additions york pa #sunroom additions charlotte nc #sunroom additions canada #sunroom additions ideas #sunroom additions #sunroom additions ideas photos"

✔ “”Gillon McClintock (gillon) Photos / 500px
Gillon McClintock
I am a hobby landscape photographer based in johannesburg”

Wrong Language (WL)

If the text is parallel data, is the target side in Modern (standardised) Irish, and the source side in Modern English? If the text is monolingual data, is the source side in Modern (standardised) Irish? (No or Unsure: annotate as WL)
Note: Untranslated named entities on either source or target side should be treated as belonging to that language (e.g. the drug name “Capnat” should be considered English text); If text is entirely named entities, it is more likely to be ‘Yes’)

✗ “Turks”
“a”

✗ “Hail Samwise the Strong!”
“Hail Samwise the Strong!”

✗ “Most of those are my dad’s.”
“em most Is Dad.”

✗ “and the greatness of His power in them.”
“agus ar mór-chumasg a iolchumhacht orra.”[4]

✗ “Deep Forest Beard Oil – ShopCaveman Home Beard Oil Deep Forest Beard Oil Grapeseed Oil | Castor Oil | Apricot Oil | Virgin Cedarwood Oil | Fir Needle Oil | Eucalyptus Oil | Spearmint Oil”

✗ “Très agréable à porter... Très agréable à porter le soir et le matin pour traîner chez soi”

✔ “Barack Obama”
“Barack Obama”

✔ “Sound policy decisions depend on evidence-based knowledge of past trends and future projections.”
“Bíonn cinntí beartais fónta ag brath ar eolas ar bhonn fianaise ar threochtaí roimhe seo agus ar réamhmheastacháin i leith na todhchaí.”

✔ “And there’s no reasoning with my Mum.”
“Níl aon réasún le Mam.”

✔ “Is there any other country in the world that’s so enthralled by celebrity?”
“An bhfuil aon tír eile ar an domhan go bhfuil an taisteal chomh lárnach sin san aistear ón déagóir go dtí an duine fásta?”

✔ “Leis na scilleanna a gheobhainn siad anseo, tá siad in ann deacrachtaí agus fadbhanna a réiteach as a stuaim féin. Tá siad ag foghlaim i dtimpeallacht nádurtha agus ag cur aithne ar a gcomhscoláirí agus a gcuid múinteoirí. Is buaicphointe na bliana é do go leor de na scoláirí!”

Incorrect Translation (X)

If the text is parallel data, is the Irish target text a direct translation of the English source text? (Yes or unsure: next question; No: annotate as X)

✗ EN: I’m returning my s10 in the morning.
GA: Ar Maidin Is Me I Dtaisce Mo Churaim10.
Lit: *In The Morning I Am In Deposit Of My Care10.*

✗ EN: Homunculus - The Other One
GA: Ceolchoirm - An Chéad Ghlúin Eile
Lit: *Concert - The Next Generation*

✗ EN: Any other suggestions of this kind?
GA: Gach bean eile de chuid an genre seo?
Lit: *Every other women of this genre?*

✔ EN: I think I hate working from home.
GA: Is fuath liom a bheith ag obair ón mbaile.
Lit: *I hate to be working from home*

✔ EN: Do not look for welcome here.
GA: Ná bí ag lorg fáilte anseo.

✔ EN: Steel and Magnesium Alloy, Nickel-Plated for Corrosion Resistance Origin.
GA: Alloy Cruach agus Maignéisiam, Neitil-Plátáilte le haghaidh Friotaíocht Creimeadh Origin.

Short text (CS)

Is this short text (Just headings (single, unrelated phrases), or five words or fewer on the Irish side)? (Yes: annotate as CS)

✗ “They specialize in many different.”
“Speisialtóireacht siad i go leor éagsúla.”

✗ “He is the great peacemaker.”
“Is í [[La Paz]] an príomhchathair.”

✗ “She spoke in broad Yorkshire in her amazement.”
“Labhair sí i Yorkshire leathana ina iontas.”

✗ “Uimhir Teagmhála: 01-6716444 Seoladh ríomhphoist: info@zebahd.ie Nasc don suíomh idirlín: zeba.ie Luan - Máirt: 9-6 Céadaoin - Déardaoin: 9-8 Aoine: 9-7 Satharn: 9-6 Is féidir do chuid gruaige a fháil gearrtha i gceann de na siopaí gruagaire is cáiliúla san Ardchathair, i measc bailte eile timpeall na tíre le gruagairí atá ar ardchaighdeán. Teagmháil Zeba - Sráid Liam Theas.”

✔ “Of course, sir.”
“Ar ndóigh, a dhuine uasail.”

✔ “-You need a diversion.”
“-Tá atreorú uait.”

✔ “I am a housewife.”
“Is bean tí mé.”

Boilerplate or Low-quality (CB)

Is this boilerplate (text that would repeat across similar webpages) or low quality text (including unnatural code-switching, unnatural phrasing, frequent unnatural formatting, and misalignments)? (Yes: annotate as CB; No, or Unsure: annotate as CC)
Note: CB includes text that is not representative of natural language, i.e. running text; this might appear as short phrases or sentence fragments. One formatting problem is not enough to disqualify text as CC; more than one could be CB

✗ “The Board of Directors shall consist of 19 directors and 11 alternates.”

[4] While this text is Irish, it does not appear to be Modern Irish (post-17th century).

“19 stiúrthóir agus 11 malartach a bheidh ar an mBord Stiúrthóiri. Thóiri.”

✗ “Tarluithe Diat Worms Tógadh Taigh BhroughtonI gCille Chùithbeirt, Alba. Breitheanna 19 Feabhra — Luigi Boccherini, cumadóir ceoil is dordveidhleadóir Iodálach 24 Feabhra — Joseph Banks, luibheolaí Sasanach (b. 1820) 13 Aibreán — Thomas Jefferson 3ú uachtarán na Stát Aontaithe Mheiriceá (b. 1826)”

✗ “Media You blocked @G_MenBasketball Are you sure you want to view these Tweets?”

“Meáin Chuir tú cosc ar @grahamgmen An bhfuil tú cinnte gur mhaith leat breathnú ar na Tweetanna seo?”

✗ “The search engine prefers to keep the article updated.”

“Is fearr leis an inneall cuardaigh an t-alt a choinneáil nuashonraithe.”

✗ “A shuiteáil ar Na botúin a bhfuil i gcónaí gearrcónaí, ach d ’ fhéadfadh scannán saintréithe anailís a chinneadh anScannán Anailís Theicniúilcourtside Cé is cosúil Sé i ndáiríre fond an fhoireann hes nach bhfuil cad ba mhaith leat cuairt a thabhairt ar antiophthalmic fachtóir scannán anailís theicniúil spóirt aficionado Nuair a d ’ iarr mé air má tá sé ina dhiaidh sin cispheile cluiche níos luaithe an Toirneach tháinig townsfolk uimhir adamhach 2 a bhí ag smaoineamh ar feadh níos lú soicind [...]”

✔ “9.Declaration on Part Three, Title XIX, of the Treaty establishingthe European Community”

“9.Dearbhú maidir le Cuid a Trí, Teideal XIX den Chonradh agbunú an Chomhphobail Eorpaigh”

✔ “Amid the growing number of challenges but also opportunities we face outside the European Union, the 73 Members of this committee contribute towards the definition of the EU’s foreign and security policy and scrutinise its implementation.”

“I measc na ndúshlán atá ag méadú, agus i measc na ndeiseanna atá os comhair Pharlaimint na hEorpa freisin, bíonn 73 Chomhalta an choiste seo ag rannchuidiú le sainiú bheartas eachtrach agus slándála AE agus déanann siad cur chun feidhme an bheartais sin a ghrinnscrúdú.”

✔ “24th April. Bhronn muintir Folláin geansaí nua arís ar iománaithe fé 9 agus fé 10. Tá Laochra Óg an bhuíoch leo go bhfuil siad chun fanúint linn i gcomhair ceithre bliain eile. Míle buíochas le Peadar as an subh a thug sé dos na páistí agus tuismitheoirí go léir. Tá súil again go mbeidh séasúr rathúil acu.”

12. Appendix B: Additional Tables and Examples

Sample	Total	Min	Mean	Max
NLLB	1,177	2	12	45
Paracrawl	2,238	3	22	89
HPLT	2,106	2	21	72
CCMatrix	1,894	2	19	57
EUbookshop	2,241	2	22	101
XLEnt	563	1	6	15
QED	1,654	1	17	108
EUconst	1,798	1	18	78
OpenSubtitles	745	1	7	23
Tatoeba	517	1	5	10
CulturaX	69,906	25	699	5,371
C4	45,709	12	462	7,379
OSCAR	6,294	16	63	632
Wikipedia	13,868	2	139	4,966
TOTAL	150,710		108	
Parallel TOTAL	14,933		15	
Monolingual TOTAL	135,777		339	

Table 9: # Tokens in Irish samples across different corpora.

Label	Example of Irish text
NL	1402A 1 402A 1-402A 14 02A 14-02A
WL	“The objects are generic.” “Author Shows Is Gen.”
X	“Emirates” “súil”

Table 10: Examples of low-quality data in manually-annotated datasets.

Label	Example of Irish text
Sexual content	“St Andrew ’s” “Brazzers físeán catagóir Inexperienced Déagóirí Seapáinis Aziatochki Téalainnis Físeán”
Personally-identifiable information	“Members with Waist : NjcgY20gfCAyNi40Ig== Giacinta Grillo* Waist: 67 cm | 26.4” Hips: 92 cm | 36.2” Edda Draghi* Hips: 94 cm | 37.0” in Ceraso*”

Table 11: Examples of harmfulness in manually-annotated datasets. *Names have been changed to prevent leaking personal information.

Label	Example of Irish text
Non-text elements User-generated content	23h00 inniu GA ina teanga oifigiúil oibre den 🇪🇺 gan mhaolú, leithscéal, cíos, cás ná cathú. An gnó céanna le tabhairt i gcríoch in 🇮🇪 san blianta beaga amach romhainn.
Code-switching User-generated content	“But you know I’m out here thinkin ’bout it.” “Ach tá a fhios agat go bhfuilim amuigh anseo thinkin ’bout é”
Non-standard dialect	“For God so loved the world, that he gave His” “Óir do ghrádhuigh Dia an saoghal chómh mór sin, go dtug sé a”

Table 12: Examples of non-standardness in manually-annotated datasets.