

# Survey of Irish Text Corpora for NLP

Abigail Walsh, Mark Andrade, Jane Adkins, Ornait O’Connell, Éanna O’Connor

ADAPT Centre, Dublin City University  
firstname.lastname@adaptcentre.ie

## 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 surveys existing monolingual and parallel Irish text corpora to evaluate the quality of the language data.

## 1 Introduction

In the early days of Large Language Model performance, 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 in this research 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.

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, data of differing domains will present with different 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 that undertaken by 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 adequate high-quality data.

In order to ascertain the status of Irish language data, we present a preliminary assessment of existing Irish language datasets, focusing on large, automatically-curated datasets. We explore the quality and source of these large resources which are likely used for generating multilingual large language models (MLLMs) that provide support to Irish.

We consider the following questions in the analysis of these datasets:

1. Does Irish language data in large multilingual datasets appear consistent with the quality and content of majority languages in the same corpus, such as English?
2. How appropriate are these datasets for use in NLP tasks?
3. What features in these large datasets (e.g. repetitive content, language ID errors, misalignments) can limit the usability of existing Irish data?

| Rank/ID | Dataset       | Size (tokens in M) |
|---------|---------------|--------------------|
| 1       | NLLB          | 250                |
| 2       | Paracrawl     | 63                 |
| 3       | HPLT          | 59                 |
| 4       | CCMatrix      | 21                 |
| 5       | EUBookshops   | 2.74               |
| 6       | XLEnt         | 0.95               |
| 7       | QED           | 0.32               |
| 8       | EUconst       | 0.15               |
| 9       | OpenSubtitles | 0.12               |
| 10      | Tatoeba       | 0.01               |
| 11      | CulturaX      | 216                |
| 12      | C4            | 103                |
| 13      | OSCAR         | 9                  |
| 14      | Wikipedia     | 8                  |

Table 1: Table of datasets surveyed with their respective size in number of Irish tokens (in Millions).

This dataset analysis serves as a foundation for strategic data collection efforts aimed at addressing critical gaps in Irish language resources, and supports future balanced and comprehensive language technology development.

## 2 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 for the collecting and sharing of high-quality Irish 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). Data selected for survey include large monolingual text corpora, and parallel text corpora, as shown in Table 1.

### 2.1 Large Monolingual Text Corpora

Monolingual text corpora represent the largest source of training data for language generation models such as LLMs, as such, the language in these corpora should represent the desired language output of the model, depending on the intended use case. We described here four corpora: CulturaX, C4, OSCAR, and Vicipéid (Irish Wikipedia).

**Vicipéid / Irish Wikipedia** Vicipéid belongs to the Wikimedia collection of online resources. It is an open source publicly available encyclopedia written and maintained by a community of volunteers following a model of open collaboration. As of 23 June 2025, it consists of 61,939 articles with 87 active editors, and currently ranks 88 out of 342 languages in terms of size (Wikipedia Contributors, 2025). The open collaboration model allows for all members of the public to create and edit articles, fostering an environment of balanced, unbiased information. However, this open collaboration means the quality of the content cannot always be assured, and this can lead to noisy or inaccurate content.

A preliminary audit of the quality of Vicipéid text found a significant variance in the quality of the data. Seven authors were identified as being particularly prolific, producing a large number of high-quality articles. As of 11 June 2025, these authors accounted for 15% of the total number of articles on Vicipéid. A version of the Vicipéid dataset is hosted on

HuggingFace (extracted 1st of November 2023), containing over 8 million tokens ([Wikimedia Foundation, 2023](#)).

**OSCAR** The OSCAR project (Open Super-large Crawled Aggregated coRpus) is an Open Source project to provide resources and datasets for machine learning and AI ([OSCAR Project, 2024](#)), specifically unannotated raw data for the pretraining of large deep learning models. The OSCAR project developed pipelines to classify and filter web data collected from CommonCrawl (<https://commoncrawl.org/>). They also focused on improving the data quality of web-based corpora and providing data for low-resource languages like Irish, to increase accessibility of resources. Pre-processing steps for this corpus are described in [Jansen et al. \(2022\)](#), and included tokenization, stopword removal, lemmatization and the replacement of emojis. The latest version of OSCAR (January 2023) contained 9 million tokens of Irish ([Godey, 2023](#)).

**Colossal Clean Crawled Corpus (C4)** The C4 dataset contains 10.4 billion documents, including nearly 500,000 documents for Irish ([Allen AI, 2023](#)). The texts here come from a variety of sources, with [Raffel et al. \(2020\)](#) stating that CommonCrawl served as the basis for the text. An examination of 20 document samples from the corpus by a fluent Irish speaker showed poor quality language data, which appeared to consist largely of English text mislabelled as Irish. The data appears to originate from varied sources, including online retailer websites and reviews, and text from the Irish radio station *Raidió na Gaeltachta*'s website and the Irish name database [ainm.ie](#).

**CulturaX** This is a massive dataset made up of 7.2 billion documents in 167 different languages, developed in partnership between the University of Oregon and Adobe Research ([Nguyen et al., 2023](#)). Pre-processing steps included language identification, URL-based filtering, metric-based cleaning (including word count, character and word ratios), threshold selection and document refinement. The Irish language data is made up of over 300,000 documents. 34% of this data is reported to overlap with the OSCAR corpus, and the remaining documents (66%) overlap with the C4 corpus ([Nguyen et al., 2023](#)). An assessment of 20 randomly selected documents by a fluent Irish speaker found the language quality varied, with many instances of code-switching. The texts featured included newspaper articles, some online shopping, and some boilerplate text.

## 2.2 Parallel Corpora

While parallel corpora can be useful in many tasks, they are particularly important for the development of machine translation engines, and are also useful for certain semantic tasks, including disambiguation. We describe here a selection of Irish-English parallel corpora that are available on the OPUS platform ([Tiedemann, 2012](#)), a freely accessible repository of parallel language data for 747 languages, including Irish.

**CCMatrix / NLLB — Parallel Subset** CCMatrix is a large dataset of high-quality web-based bitexts for training translation models developed by Meta AI. With more than 4.5 billion parallel sentences in 576 language pairs pulled from snapshots of the CommonCrawl public dataset, CCMatrix is more than 50 times larger than the WikiMatrix corpus previously shared by the NLLB research team. There are over 1 million sentences in the English-Irish dataset in CCMatrix. A sample of documents from this dataset displayed on the OPUS platform mostly includes religious documents, from the Old Testament and the Quran.

The NLLB parallel English-Irish dataset is made up of more than 22.5 million sentences. The original NLLB data was collected using a modified version of the bitext mining approach used for WikiMatrix ([Schwenk et al., 2019](#)), assuming that the translation of one sentence could be found anywhere on CommonCrawl. Massively parallel processing and the FAISS library for fast similarity searches were employed to address the significant computational challenges posed by comparing billions of sentences to determine which ones are mutual translations. A random selection of 20 samples, each approximately 10 sentences long, showed the data originated from varied sources, including legal and governmental documents.

**Paracrawl** Paracrawl is an ongoing large-scale web-crawling project, available through the OPUS repository. The project used web crawling tools to gather parallel datasets across European languages. This process enables the creation of a large dataset, but the quality is not assured, and it can contain noise, such as non-post-edited MT output (Defauw et al., 2019).

According to OPUS, it has a 9% overlap with HPLT and a 3% overlap with NLLB (Tiedemann, 2012). A random sample of 100 sentences was assessed by a fluent Irish speaker, and found text originating from varied domains, including legal, news, technical, social media, and adult content.

**HPLT** The HPLT (High-Performance Language Technologies) corpus includes monolingual and bilingual corpora sourced from CommonCrawl and previously unused web crawls from the Internet Archive (de Gibert et al., 2024). The language of each document was identified using FastSpell. Both Monofixer and Monocleaner were used to process the text. Deduplication was also performed.

The latest version (v2) contains 803 million sentence fragments, or 17 billion tokens. The English-Irish dataset contains nearly 2.7 million sentences. Surveying the sample texts on the OPUS platform, the data appears to originate from a variety of sources, and include several instances of code-switching. The samples included a sentence in Zulu which was mislabeled as Irish. OPUS reports a 7% overlap with the Paracrawl corpus, and a 5% overlap with NLLB (Tiedemann, 2012).

**QED** The QED Corpus (QCRI Educational Domain Corpus) is a collection of subtitles from educational content that contains approximately 371 million tokens. The English-Irish dataset contains over 18,000 sentences. It has a 1% overlap with the NLLB dataset (Tiedemann, 2012). According to Abdelali et al. (2014), rather than a standard processing system, the QED Corpus was filtered to remove incomplete subtitles and those that were labeled incorrectly. An assessment of a random sample of 100 sentences by two fluent Irish speakers found the text to be of varied quality, with high instances of code-switching.

**XLent** 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). From the samples shown on OPUS, there are alignment issues with almost all the texts. It is not possible to judge the grammatical quality as the data is segmented by named entity. A random selection of 20 samples was assessed by a fluent Irish speaker for both quality and alignment, and appeared to be of generally low quality and poor alignment.

**OpenSubtitles** The OpenSubtitles corpus consists of 2.6 billion sentences in 60 different languages from film subtitles (Lison & Tiedemann, 2016). The English-Irish data consists of 20,035 sentences. A random selection of 20 samples assessed by a fluent Irish speaker were deemed to be of largely good quality, and appeared to originate from film subtitles.

**EU Bookshops** The Publications Office of the European Union (EUBookshops) contains all official publications made by the EU. These documents are stored in a corpus on OPUS, made up of 135,785 files. There are 246 Irish language datasets comprising 4 million tokens, with parallel texts in 34 different languages. These range from just 5 pairs with Albanian and Swahili to over 133,000 sentences pairs with English. According to Tiedemann (2012), there is a 7% overlap with ParaCrawl, 5% with NLLB and HLPT, as well as a 2% overlap with EUConst. As with DGT, the documents here regard European Union matters, but are of a slightly less formal nature.

**EUConst** The EUConst dataset is made up of parallel documents of the European Union's constitution. OPUS shows it has a 31% overlap with EUBookshops, as it contains the constitution within it. It also has a 24% overlap with Paracrawl and 9% overlap with HPLT (Tiedemann, 2012).

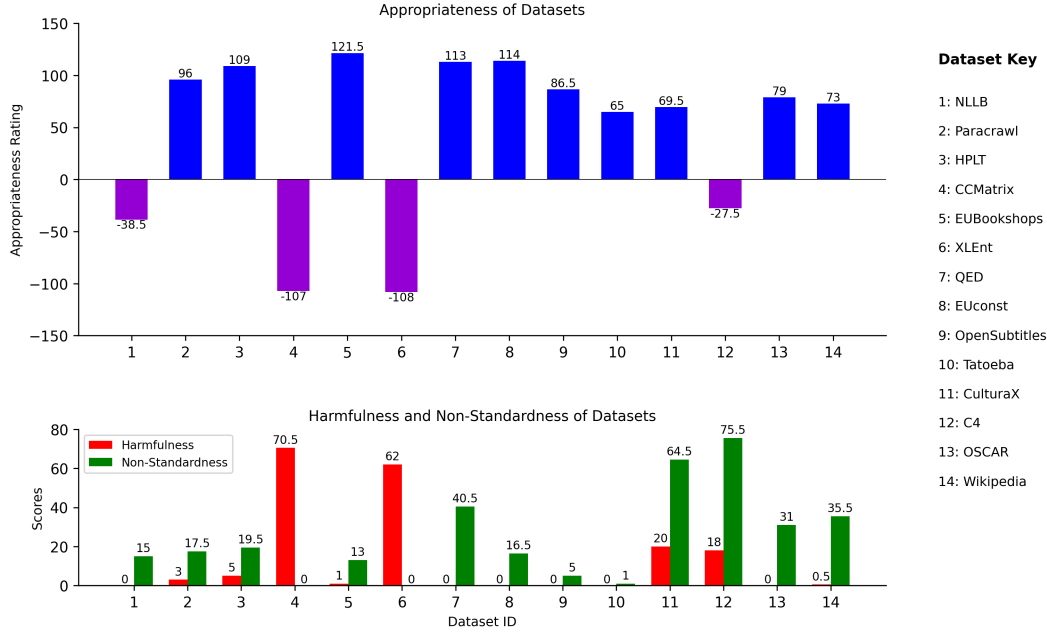


Figure 1: Appropriateness, harmfulness and non-standardness of datasets

**Tatoeba** Tatoeba is a small dataset of short learner-friendly sentences with simple vocabulary and grammar, available on OPUS. There are just over 2,000 sentences in the English—Irish parallel dataset. It has a 2% overlap with NLLB and a 1% overlap with Paracrawl (Tiedemann, 2012).

### 3 Quality Annotation

Following preliminary analysis, 100 samples were selected randomly from the fourteen datasets above for quality annotation. Samples from parallel data include an English sentence and the equivalent Irish translation. Samples from monolingual datasets were extracted from HuggingFace (HuggingFace, 2025) using their integrated DuckDB service, and formatted as running text belonging to single document. Samples ranged in length from a single sentence to almost one hundred sentences, although typically samples were between 10-20 sentences. The data was annotated across three dimensions: appropriateness, harmfulness, and non-standard text elements.

**Appropriateness** was determined using the six quality annotation labels outlined by the dataset quality audit performed by Kreutzer et al. (2022). Table 3 explains how these labels were applied. **Harmfulness** was identified through five labels, as informed by Chehbouni et al. (2025): *racist or offensive language*, *illegal or antisocial behaviour*, *sexual content*, *personally identifiable information*, and *other*. **Non-standardness** was annotated using four labels, based on preliminary audits of the Irish data: *non-standard dialect*, *code-switching*, *user-generated content*, and *non-textual elements*. Both harmfulness and non-standardness labels were applied to any instances detected in the samples, and one sample could be annotated with multiple labels. However, each instance was only counted once per sample.

Following the annotation task, a score for each dimension was calculated using the labels assigned to the data. An **appropriateness score** was calculated by assigning a value to each labelled sample as follows: Correct translation, natural sentence (CC)=2, Correct translation, short (CS)=1, Correct translation, boilerplate or low-quality (CB)=0, Incorrect translation, but both correct languages (X)=-1, Source OR target wrong language, but both still linguistic content (WL)=-2, and Not a language: at least one of source and target are not linguistic content (NL)=-3. The scores for each sample were then added up for a total appropriateness

| Dimension                  | $\alpha$ -coarse | $\kappa$ -coarse | $\alpha$ -fine | $\kappa$ -fine |
|----------------------------|------------------|------------------|----------------|----------------|
| Alignment (A and B)        | 0.48             | 0.49             | 0.7            | 0.7            |
| Harmfulness (A and B)      | 0.87             | 0.88             | 0.87           | 0.88           |
| Non-standardness (A and B) | 0.48             | 0.47             | 0.41           | 0.44           |
| Alignment (C and D)        | 1.0              | 1.0              | 0.57           | 0.58           |
| Harmfulness (C and D)      | 0.63             | 0.67             | 0.63           | 0.67           |
| Non-standardness (C and D) | 0.03             | 0.16             | 0.05           | 0.12           |

Table 2: Coarse-grained and Fine-grained average IAA scores for Annotators A and B, and Annotators C and D.  $\alpha$  stands for Krippendorff’s Alpha and  $\kappa$  stands for Cohen’s Kappa.

score for each dataset. 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.<sup>1</sup> The results of this annotation are explored in Section 3.2 below.

### 3.1 Inter Annotator Agreement

Four annotators completed the annotation task. Applying the Common European Framework of Reference for Languages (CEFR) Self Assessment Grid for reading comprehension<sup>2</sup>, Annotator A had a C1 level of comprehension, Annotators B and C had a B2 level of comprehension, and Annotator D, a professional Irish translator, had a C2 level of comprehension.

The parallel data was annotated by Annotators A and B, and the monolingual data was annotated by Annotators C and D. Each annotator annotated 40 samples individually, and 20 samples were doubly annotated, for a total of 100 samples per dataset. Inter Annotator Agreement (IAA) was calculated using Krippendorff’s alpha and Cohen’s kappa for the doubly annotated portion of each dataset, by comparing the labels applied to each sample. Fine-grained agreement required complete agreement of the labels applied to each sample, including potentially multiple labels assigned in harmfulness and non-standardness dimensions. Coarse-grained scores were calculated using broad agreement level labels: samples annotated with CC, CS, or CB are labelled as ‘correct’, and samples annotated with X, WL or NL are labelled as ‘incorrect’. Harmfulness and Non-standardness were said to agree where both annotators detected an instance of either dimension. Fine-grained scores were calculated based on agreement of exact label used for each dimension. The results of agreement scores averaged across all datasets are shown in Table 2.

In general, IAA scores show positive agreement, particularly across the harmfulness dimension. Non-standardness agreement was low for Annotators C and D, indicating more annotation guidelines would be helpful, particularly for identifying instances of non-standard text, which appears to have been applied frequently in cases of ungrammatical text. Examining IAA scores of individual datasets shows 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.

### 3.2 Results

The results of the annotation scores are shown in Figure 1. As can be seen in the graphs, several datasets (NLLB, CCMatrix, XLEnt, and C4) resulted in negative appropriateness scores when the scores from annotation labels were measured. The text from these datasets tended to include non-text elements such as markup language or metadata tags, and the text was often not in Irish. It is not clear whether pre-processing steps were applied to the Irish sample of these datasets, or if pre-processing steps failed due to lack of high-quality tools trained to handle Irish text. Unsurprisingly, EU Bookshops and EUConst, two

<sup>1</sup>The scores for the doubly-annotated portions that were annotated by two annotators were averaged across both annotators.

<sup>2</sup><https://www.coe.int/en/web/common-european-framework-reference-languages/table-2-cefr-3.3-common-reference-levels-self-assessment-grid>

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

collections of data originating EU publications, show very high appropriateness scores, and low harmfulness scores.

Parallel datasets tended to score better on appropriateness scores than monolingual datasets, perhaps due to the samples being much smaller. This is also reflected in non-standardness scores, which are clearly higher for monolingual datasets, unsurprising given that the larger monolingual samples were more likely to include an instance of non-standard text, such as symbols or code-switching. Somewhat surprisingly, the harmfulness scores for the monolingual datasets were not as high as scores for CCMatrix or XLEnt, two parallel datasets, which both featured a high-degree of sexual content, likely machine-translated. The extremely low appropriateness scores for these two datasets additionally reflects the Irish text, which was almost entirely unrelated to their English source sentences.

## 4 Conclusion

Based on our preliminary analysis of monolingual Irish corpora, Vicipéid, OSCAR, and CulturaX were all of similar quality, while the quality of text in C4 was generally poor. Parallel English-Irish corpora such as EU Bookshops, EUConst, QED, and HPLT all demonstrate high-quality data, with low scores in harmfulness, while NLLB, CCMatrix and XLEnt datasets present serious linguistic, structural reliability, and/or harmfulness concerns, making them unsuitable for use in the construction of Irish NLP pipelines, and particularly unsuitable for the training of Irish MT models.

The above analysis represents a limited investigation into several monolingual and parallel Irish text corpora. However, with even this limited analysis, we can observe quality issues and language features present in these corpora that may limit their usability for training NLP models—including mislabelling Irish language, misalignments, machine-generated text, code-switching, grammatical errors, and duplicated text. Further analysis is desirable to understand the impact of these quality issues or language features on downstream tasks, such as machine translation.

## References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 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)*, pp.

- 1044–1054, Reykjavik, Iceland, 2014. European Language Resources Association (ELRA). URL [http://www.lrec-conf.org/proceedings/lrec2014/pdf/877\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2014/pdf/877_Paper.pdf).
- Allen AI. Allen AI: Irish dataset, 2023. URL <https://huggingface.co/datasets/allenai/c4/viewer/ga>. Accessed: 25 June 2025.
- James Barry, Joachim Wagner, Lauren Cassidy, Alan Cowap, Teresa Lynn, Abigail Walsh, Mícheál J. Ó Meachair, and Jennifer Foster. gaBERT — an Irish language model. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4774–4788, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.511/>.
- Khaoula Chehbouni, Jonathan Colaço Carr, Yash More, Jackie CK Cheung, and Golnoosh Farnadi. Beyond the safety bundle: Auditing the helpful and harmless dataset. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 11895–11925, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.596. URL <https://aclanthology.org/2025.naacl-long.596/>.
- Teresa Clifford, Abigail Walsh, Brian Davis, and Mícheál J. Ó Meachair. Gaeilge bhriste ó shamhlacha cliste: How clever are LLMs when translating Irish text? In Brian Davis, Theodorus Fransen, Elaine Uí Dhonnchadha, and Abigail Walsh (eds.), *Proceedings of the 5th Celtic Language Technology Workshop*, pp. 46–51, Abu Dhabi [Virtual Workshop], January 2025. International Committee on Computational Linguistics. URL <https://aclanthology.org/2025.cltw-1.5/>.
- Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoyong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo<sup>1</sup>, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. A new massive multilingual dataset for high-performance language technologies. *arXiv preprint*, 2024.
- Arne Defauw, Sara Szoc, Tom Vanallemeersch, Anna Bardadym, Joris Brabers, Frederic Everaert, Kim Scholte, Koen Van Winkel, and Joachim Van den Bogaert. Developing a neural machine translation system for Irish. In Alina Karakanta, Atul Kr. Ojha, Chao-Hong Liu, Jonathan Washington, Nathaniel Oco, Surafel Melaku Lakew, Valentin Malykh, and Xiaobing Zhao (eds.), *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pp. 32–38, Dublin, Ireland, August 2019. European Association for Machine Translation. URL <https://aclanthology.org/W19-6806/>.
- Meghan Dowling, Teresa Lynn, and Andy Way. Leveraging backtranslation to improve machine translation for Gaelic languages. In Teresa Lynn, Delyth Prys, Colin Batchelor, and Francis Tyers (eds.), *Proceedings of the Celtic Language Technology Workshop*, pp. 58–62, Dublin, Ireland, August 2019. European Association for Machine Translation. URL <https://aclanthology.org/W19-6908/>.
- Ahmed El-Kishky, Adithya Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. XLEnt: mining a large cross-lingual entity dataset with lexical-semantic-phonetic word alignment. *arXiv preprint arXiv:2104.08597*, 2021. URL <https://arxiv.org/pdf/2104.08597>.
- Nathan Godey. OSCAR Small: Unshuffled and deduplicated Irish dataset, 2023. URL [https://huggingface.co/datasets/nthngdy/oscar-small/viewer/unshuffled\\_deduplicated\\_ga](https://huggingface.co/datasets/nthngdy/oscar-small/viewer/unshuffled_deduplicated_ga). Accessed: 25 June 2025.
- Christoph Gröger. There is no AI without data. *Communications of the ACM*, 11:98–108, 2021.
- HuggingFace. Huggingface, 2025. URL <https://huggingface.co/>.

- Tim Jansen, Yangling Tong, Victoria Zevallos, and Pedro Ortiz Suarez. Perplexed by quality: A perplexity-based method for adult and harmful content detection in multilingual heterogeneous web data. *arXiv preprint arXiv:2212.10440*, 2022. URL <https://arxiv.org/abs/2212.10440>.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL <https://aclanthology.org/2020.acl-main.560/>.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL <https://arxiv.org/abs/2001.08361>.
- Amr Keleg, Matthias Lindemann, Danyang Liu, Wanqiu Long, and Bonnie L. Webber. Automatically discarding straplines to improve data quality for abstractive news summarization. In Tatiana Shavrina, Vladislav Mikhailov, Valentin Malykh, Ekaterina Artemova, Oleg Serikov, and Vitaly Protasov (eds.), *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pp. 42–51, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.nlppower-1.5. URL <https://aclanthology.org/2022.nlppower-1.5/>.
- 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, Iroko 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 Baruwu, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022. doi: 10.1162/tacl.a.00447. URL <https://aclanthology.org/2022.tacl-1.4/>.
- Seamus Lankford, Haithem Afli, and Andy Way. Machine translation in the covid domain: an English-Irish case study for LoResMT 2021. In John Ortega, Atul Kr. Ojha, Katharina Kann, and Chao-Hong Liu (eds.), *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pp. 144–150, Virtual, August 2021. Association for Machine Translation in the Americas. URL <https://aclanthology.org/2021.mtsummit-loresmt.15/>.
- Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report, 2023. URL <https://arxiv.org/abs/2309.05463>.
- Pierre Lison and Jörg Tiedemann. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pp. 923–929, 2016. URL <https://aclanthology.org/L16-1147.pdf>.
- Teresa Lynn. Report on the Irish language. <https://european-language-equality.eu/deliverables/>, 2022. Technical Report D1.20, European Language Equality Project.
- Yida Mu, Mali Jin, Xingyi Song, and Nikolaos Aletras. Enhancing data quality through simple de-duplication: Navigating responsible computational social science research. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12477–12492, Miami,

- Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.694. URL <https://aclanthology.org/2024.emnlp-main.694/>.
- Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*, 2023. URL <https://arxiv.org/abs/2309.09400>.
- Mícheál Ó Meachair, Úna Bhreathnach, and Gearóid Ó Cleirín. Introducing the national corpus of Irish project. In Theodorus Fransen, William Lamb, and Delyth Prys (eds.), *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, pp. 99–103, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.cltw-1.14/>.
- Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Toluwalase Owodunni, Odunayo Ogundepo, David Ifeoluwa Adelani, and Jimmy Lin. Better quality pre-training data and t5 models for African languages. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 158–168, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.11. URL <https://aclanthology.org/2023.emnlp-main.11/>.
- OSCAR Project. OSCAR: Open Source Project on Multilingual Resources for Machine Learning, 2024. URL <https://oscar-project.org/>. Accessed: 25 June 2025.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michał Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL <https://www.jmlr.org/papers/volume21/20-074/20-074.pdf>.
- Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. CCMatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*, 2019. URL <https://arxiv.org/abs/1911.04944>.
- Lilli Smal, Andrea Lössch, Josef van Genabith, Maria Giagkou, Thierry Declerck, and Stephan Busemann. Language data sharing in European public services – overcoming obstacles and creating sustainable data sharing infrastructures. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 3443–3448, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL <https://aclanthology.org/2020.lrec-1.422/>.
- Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pp. 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL <https://aclanthology.org/L12-1246/>.
- Khanh-Tung Tran, Barry O’Sullivan, and Hoang Nguyen. Irish-based large language model with extreme low-resource settings in machine translation. In Atul Kr. Ojha, Chao-hong Liu, Ekaterina Vylomova, Flammie Pirinen, Jade Abbott, Jonathan Washington, Nathaniel Oco, Valentin Malykh, Varvara Logacheva, and Xiaobing Zhao (eds.), *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pp. 193–202, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.loresmt-1.20. URL <https://aclanthology.org/2024.loresmt-1.20/>.
- Khanh-Tung Tran, Barry O’Sullivan, and Hoang D. Nguyen. Uccix: Irish-excellence large language model, 2024b. URL <https://arxiv.org/abs/2405.13010>.

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

Wikimedia Foundation. Wikipedia language edition: Irish (ga), Dump from 1 november 2023, 2023. URL [https://huggingface.co/datasets/wikimedia/wikipedia/viewer/20231101.ga?views%5B%5D=\\_20231101ga](https://huggingface.co/datasets/wikimedia/wikipedia/viewer/20231101.ga?views%5B%5D=_20231101ga). Accessed: 25 June 2025.

Wikipedia Contributors. List of Wikipedias, 2025. URL [https://en.wikipedia.org/wiki/List\\_of\\_Wikipedias](https://en.wikipedia.org/wiki/List_of_Wikipedias). Accessed: 25 June 2025.