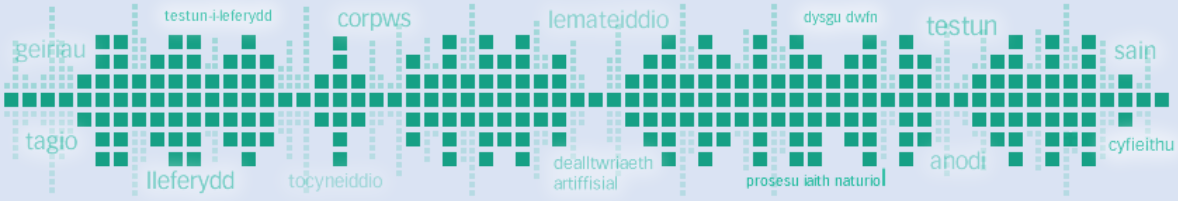
Language and Technology in Wales: Volume II

Watkins, Gareth; Prys, Delyth; Prys, Gruff; Jones, Dewi; Cooper, Sarah; Williams, Meinir; Vangberg, Preben; Ghazzali, Stefano; Gruffydd, Ianto; Farhat, Leena; Grobol, Loïc; Jouitteau , Mélanie; Morris, Jonathan; Ezeani , Ignatius; Young , Katharine; Davies, Lynne; El-Haj , Mahmoud; Knight , Dawn; Jarvis , Colin; Barnes, Emily

Published: 01/11/2024

Publisher's PDF, also known as Version of record



Language and Technology in Wales: Volume II

Editor: Gareth Watkins

PRIFYSGOL BANGOR UNIVERSITY

Ariennir gan Lywodraeth Cymru
Funded by Welsh Government

This ebook was first published in 2024 by

Bangor University, College Road, Bangor, Gwynedd LL57 2DG

www.bangor.ac.uk

International Book Number (ebook):

ISBN 978-1 84220-207-4.

Design and proofreading assistance from Prof. Delyth Prys and Stefano Ghazzali. This book is also available in Welsh under the title *Iaith a Thechnoleg yng Nghymru: Cyfrol II*, ISBN number 978-1 84220-208-1.

Language and Technology in Wales: Volume 2

Editor:
Gareth Watkins

Contributors:
Mélanie Jouitteau UNIVERSITE DE PAU ET DES PAYS DE L'ADOUR, AND UNIVERSITÉ BORDEAUX-MONTAIGNE

Loïc Grobol UNIVERSITÉ PARIS NANTERRE

Jonathan Morris CARDIFF UNIVERSITY

Ignatius Ezeani LANCASTER UNIVERSITY

Ianto Gruffydd BANGOR UNIVERSITY

Katharine Young CARDIFF UNIVERSITY

Lynne Davies CARDIFF UNIVERSITY

Mahmoud El-Haj LANCASTER UNIVERSITY

Dawn Knight CARDIFF UNIVERSITY

Preben Vangberg BANGOR UNIVERSITY

Leena Sarah Farhat BANGOR UNIVERSITY

Colin Jarvis OPENAI

Dewi Bryn Jones BANGOR UNIVERSITY

Gruffudd Prys BANGOR UNIVERSITY

Emily Barnes TRINTY COLLEGE DUBLIN

Meinir Williams BANGOR UNIVERSITY

Sarah Cooper BANGOR UNIVERSITY

Stefano Ghazzali BANGOR UNIVERSITY

Delyth Prys BANGOR UNIVERSITY

Contents

Preface

PROFESSOR EMERITUS DELYTH PRYS, FORMER HEAD OF THE LANGUAGE TECHNOLOGIES UNIT, CANOLFAN BEDWYR, BANGOR UNIVERSITY

When I joined Bangor University in 1993 there was very little understanding of the huge role computers, the world-wide web and digital communications would have on everyone's lives within thirty years. English was the dominant language of these media, to be gradually joined by other large global languages. It was a very different story for small, less-resourced languages, including endangered and minoritized languages such as Welsh. To them, the advent of these new technologies and digital media were a threat, endangering their ability to survive and thrive. The new technologies represented power and excitement, economic opportunities and a prosperous future, making the languages that could not access them appear old fashioned and impoverished, and therefore less likely to be transmitted to the next generation.

It was against this background that a small group of us at Bangor University decided to start developing language technologies for Welsh, partnering as far as possible with industry and the volunteer sector so that the wider community could benefit from our research. Our aim was to enable the Welsh language to obtain the necessary resources to thrive in the digital world, and that is now beginning to be realised in areas such as speech technologies, text technologies, and machine translation.

But this is a fast-moving world, and even as we were publishing the first volume of Language and Technology in Wales in 2021, few of us foresaw the forthcoming revolution that would soon arrive in the wake of generative AI such as ChatGPT. This technology can work for many languages given enough text or speech data to train it, and 'enough' here means vast sums of data, involving at least a billion parameters according to latest estimates, something that very small languages struggle to achieve.

Fortunately there are ways of working together across different linguistic communities to learn from each other and develop new techniques to achieve our aim. This is happening across Europe and the rest of the world, and closer to home, researchers working on Welsh and other Celtic languages have begun to collaborate, since there are many social, political and linguistic reasons for sharing our experiences and techniques.

The results of some of this collaboration may be seen in the present volume, and more is yet to come. Support from the European Union, the Welsh Government and other bodies have helped Celtic language technologies mature, and I look forward to seeing future developments.

I wish to thank all my colleagues at Bangor University, across Wales and beyond, and especially in Celtic language technologies for all their support and inspiration during the last thirty years.

Introduction

GARETH WATKINS – BANGOR UNIVERSITY

We publish this volume, the second in the Language and Technology in Wales series, during an incredibly exciting period in the world of Language Technology (LT). Artificial Intelligence (AI) is not a new concept,[1] however over the past year it has not only ignited the interest of academics, but has become part of the everyday lives of the general public. Recently, AI has been ever present in the newspapers, tabloids and broadsheets alike. AI has been made more accessible than ever before. Those with access to the internet are able to make use of AI through chatbots such as Microsoft's Copilot and OpenAI's ChatGPT. Uses for AI are diverse and many, and not limited specifically to LT or chatbots. For instance, AI is being used to improve crop production,[2] to identify counterfeit paintings on e-bay,[3] and to upscale and improve television images.[4]

AI then has the potential to do great good, but if this technology is not made available to minoritized languages it could have a devastating effect in terms of the perceived status and the use of those languages. AI needs to be made available for minoritized languages, then, in order that there are no linguistic barriers to using said technology, and in order that minoritized language speakers don't turn away from their language to use AI.

Effects on language use and status aside, this technology still needs further development, and that in areas which affect minoritized and majority languages alike. It can suffer from hallucinations, where the system:

> "perceives patterns or objects that are non-existent or imperceptible to human observers, creating outputs that are nonsensical or altogether inaccurate." [1]

As has been widely reported, the energy used, and therefore environmental impact of training and using AI models, is worryingly high.[5] Even a normal everyday single use (or inference) of the technology has an alarmingly high energy cost and carbon footprint. In discussing the results of their research, Luccioni et al. [2] note that:

> "the most efficient text generation model uses as much energy as 16% of a full smartphone charge for 1,000 inferences, whereas the least efficient image generation model uses as much energy as 950 smartphone charges (11.49 kWh), or nearly 1 charge per image generation."

Some are concerned about the impact of AI on jobs and the economy, indeed Georgieva [3] claims that "AI will affect almost 40 percent of jobs around the world, replacing some and complementing others", and that this figure may increase to 60 percent in advanced economies.

Clearly it is true that there remains work to be done on mitigating these issues.

[1] Those who are interested in AI's long history should read Rockwell Anyoha's informative article available at https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

[2] See for instance https://www.dmu.ac.uk/about-dmu/news/2024/april/dmu-using-ai-to-aid-crop-production-and-help-farmers-boost-income.aspx

[3] As reported by the Guardian: https://www.theguardian.com/artanddesign/article/2024/may/08/fake-monet-and-renoir-on-ebay-among-counterfeits-identified-using-ai

[4] See for instance https://www.samsung.com/ca/support/tv-audio-video/how-to-use-the-intelligent-mode-of-samsung-qled-tvs/

[5] See for instance https://www.theguardian.com/technology/2024/mar/07/ai-climate-change-energy-disinformation-report

It is also true that pre-AI boom technology remains relevant. Research continues apace in the fields of Natural Language Processing (NLP) or Corpus Linguistics, for instance. It is also important that outputs of previous research is maintained so that those outputs remain relevant and useful. Despite the recent advancements in AI, old challenges remain, most notably, perhaps, in respect of availability and collection of minoritized language data.

This volume ties together many diverse aspects of LT in the context of the Welsh language and other minoritized languages. As with the previous volume, this volume is a contribution to the development of the field in Wales, and is published bilingually under an open licence (specifically CC-BY 4.0) so that others can use it under the terms of that licence. It will also be added to Bangor's permissively licenced corpus, contributing more data which can be used in the development of LT.

Much has changed since the publication of Language and Technology in Wales Volume 1. Much is likely to change in the coming years. Yet much has and will stay the same. It is anticipated that this volume will help educate and inspire future and current researchers, and through them, contribute to future LT development.

REFERENCES

[1] IBM. 2024. What are AI hallucinations? Retrieved from https://www.ibm.com/topics/ai-hallucinations

[2] Alexandra Sasha Luccioni, Yacine Jernite and Emma Strubell. 2023. Power Hungry Processing: Watts Driving the Cost of AI Deployment? Nov. 28 2023. arXiv:2311.16863v1.

[3] Kristalina Georgieva. 2024. AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity. Retrieved from https://www.imf.org/en/Blogs/Articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity

Make Wikigrammars!

MÉLANIE JOUITTEAU

l'Université Bordeaux Montaigne, France and l'Université de Pau et des Pays de l'Adour, France

LOÏC GROBOL

Université Paris Nanterre, France

This chapter presents an evaluation from a natural language processing (NLP) perspective on the concept of wikigrammars, using the Breton ARBRES wikigrammar as a case study. It explores the utilisation of a wiki-based platform for documenting the syntactic diversity of a low-resource Celtic language, with an interactive component aimed at community engagement. It constitutes a comprehensive, annotated corpus that supports both theoretical linguistics and NLP. We advocate for the adoption of such platforms by communities speaking minoritized languages, arguing that they provide corpora with rich syntactic, orthographic, and stylistic diversity. The artificially selected diversity of wikigrammars may mitigate the scarcity of extensive, freely available corpora in low-resource language contexts.

Keywords: Wikigrammars, NLP, Breton, Corpus

1 INTRODUCTION

A wikigrammar is a wiki-based platform describing a language that is open to contributions and discussions, and whose examples are annotated and automatically retrievable.

Wikigrammars provide a very specific type of resource for language technology development: a corpus that is by definition a concentrate of linguistic diversity. In this article, we present some features of the ARBRES wikigrammar of the dialects of Breton [1], a low-resource Celtic language. We recommend the adoption of this solution by communities of minoritized languages to foster the development of their digital resource ecosystem for language technologies.

2 LINGUISTIC DIVERSITY BY DESIGN

The primary goal of ARBRES is to provide a comprehensive description of Breton, capturing its diversity, complexity and regular features, accessible in its online and written forms for the speech community. Such a grammar should not only mention and describe the most common structures, but also exceptions and infrequent phenomena. The statistical distributions of words and structure are therefore much more diverse than they would be in a random sample of the language of a similar size. Another effect contributes to the diversity of the data: for copyright reasons, the author could only take a modest percentage of the sentences for each published corpus. The effect is a widening of the variety of sources for printed free corpora (literature, newspaper articles, novels, songs, poems, collections of popular expressions, political leaflets, town hall presentation sites, posts on social networks, etc.).

A second goal of ARBRES is to be a documented resource for ongoing debates in theoretical linguistics. In that way, it is akin to a regular research notebook for its main author. This has also had an influence on the resulting data: it includes the somewhat artificial sentences typical of grammars and research papers. However, these are significantly outweighed by more natural examples. This source of data includes minimal pairs and negative evidence.

Another major source of examples is fieldwork elicitation data, where native speakers have been subjected to protocols of questions, translations or descriptive tasks of images. The raw results of the elicitations are posted online and feed the grammar. These protocols also include tasks of judgments of grammaticality of sentences, resulting in ungrammatical example that serve as contrastive negative evidence. This source of data thus also includes minimal pairs and negative evidence.

2.1 Dialectal and historical diversity

ARBRES is a grammar of dialects and has by design a high dialectal diversity. It is a descriptive grammar, where standard Breton is merely one dialect among others. The dialectal spectrum is therefore quite broad, with the notable exception of the Gwenedeg dialect, which is linguistically the furthest from the others, and is currently underrepresented in ARBRES. Its analysis requires expertise that the main editor is sometimes lacking, and as a result less data represents this dialect.

Aside from this particular caveat, we can consider that quantitatively, rare dialectal features are over-represented in the data. Indeed, common linguistic features are only illustrated with a few examples for each major dialect. On the contrary, to be able to precisely describe a rare feature, its dialect distribution and the parameters of their context of appearance, each existing occurrence will be carefully integrated. Rare features are also more likely to be the subject of thematic elicitation research, which provides more data where they occur. For the same purpose of describing the variation, the forms of different styles will co-exist within the corpus, with a quantitative over-representation of this variation compared to any single corpus. In this sense, while ARBRES is questionable for quantitative studies, it is very well suited for qualitative ones.

Finally, while ARBRES is not strictly speaking a diachronic work, it still includes data from Middle Breton to 21st century Breton. The presence of written corpus data from these periods implies, especially for the twentieth century, the presence of several competing orthographic systems. The source data has not been altered, and examples appear in their original printing spellings. The result is a multi-orthographic corpus.

2.2 Size

The ARBRES website has been developed since 2007, and started having an online presence in 2009. In the last five years, it has received more than 100 human visits per day. As of February 2024, the site consists of 10,238 pages, including 4,804 pages of content, 19 pages of presentations, and a number of redirection pages. The content pages consist of 3,094 articles on elements of Breton grammar and 325 sheets which each provide an explanation of a linguistic term or concept. Overall, this amounts to about 15,000 original Breton sentences, glossed and translated into French, coming from 1,208 research works on the Breton language (books, dictionaries, research articles, data collection blogs), 493 corpus references produced by native speakers (mostly in written forms: novels, newspaper articles, songs) and 44 elicitation sessions with native speakers.

3 A DATA SOURCE FOR NLP

3.1 Morphosyntactic annotations

The examples are provided under the form of wikitables, tables in the markup language of the MediaWiki software [2] that powers ARBRES. Each of these tables provide for a single sentence alignment of the original word forms and their glosses, the global translations of the sentence, the name of the dialectal variety, and the reference of the

source. Each word form gloss is connected via a hyperlink to a dedicated page, including at least its standard lemma and its grammatical category. Given the variety of possible spellings, this allows a high consistency in the data without being detrimental to diversity.

This system also makes it possible to reach all the word forms for a given lemma, which is crucial in this Celtic language, where inflexions are not only suffixes, but also modifications of the initial consonant depending on the syntactic context (consonant mutations). The lemma *krokodil* can thus be automatically linked to its occurrences in *krokodil Maia* (the crocodile of Maia), *ar c'hrokodil* (the crocodile), *ar c'hrokodiled* (the crocodiles) and *war grokodileta* (about to look for crocodiles), all these occurrences pointing to the page for the lemma *krokodil*. Conversely, disambiguation pages provide clickable lists of morphemes and words with more than a single meaning.

From a language technology point of view, this means that the glosses on ARBRES are already a morphosyntactically annotated corpus: a set of sentences, with lemmas and part-of-speech tags for every word and additional morphological features. It also makes it a very good seed for growing [3]. For additional details on the recoverable grammatical annotations, see Jouitteau and Bideault [4].

3.2 Parallel data

All the glosses also include translations in French, either sourced from their original publication or provided by the author, but in all cases by fluent speakers of Breton. While these translations were originally provided merely to help non-speakers make sense of the source material, they can also be seen as a parallel corpus of sentences.

This corpus is of a modest size, but it is of a very high quality and has a much larger diversity than a random sample of an equivalent size would have. Its quality simply comes from the origin of the data: all sentences have been manually selected, translated by fluent speakers, and validated carefully to ensure their relevance as illustrations of linguistic phenomena. The high diversity is ensured by the function of the glosses: since they are meant to illustrate as many linguistic phenomena as possible while taking dialectal variations into account, rare phenomena will be over-represented relatively to their organic occurrences.

Currently ongoing experiments in developing machine translation system using an early extraction of this data (around 5,000 deduplicated sentences, after removal of negative data and instances of failed data extraction) tend to confirm that these characteristics make ARBRES a very valuable dataset. Indeed, its inclusion in the training data of off-the-shelf systems result in performance gains that are comparable to those obtained with an order of magnitude more data [5].

3.3 Cost estimates

Using wikigrammars as sources of linguistic data is expensive in that it requires one or more people trained in the language, with a certain dialectal flexibility, and a social network suitable for reaching speakers of different linguistic profiles. It also requires technical support to design and maintain the website, and ensure its accessibility. Data extraction also requires qualified workers. The more laborious task is the extensive coding of examples for their appropriate presentation within wikitables. The complexity of this task is evolving fast, due to improvements in natural language generation. For the Breton wikigrammar, it took 15 years for a single annotator, working about half-time on it, to barely reach annotation of 15,000 sentences.

Chatbots now enable the automation of a significant portion of the annotation work. For instance, as of 2024, with a suitably detailed prompt providing seven examples of structured data, Chat GPT 3.5 is capable of distributing tokens across tables, aligning glosses, encoding a large fraction of clickable links, offering translations (inaccurate,

but correctly aligned), and properly organising source references. Manual interventions by a language expert are still indispensable, but they have been significantly simplified, to the point where a single person can easily enter 300 examples per month. ChatGPT 4 enhances this process even further with superior translations. Of course, this last capability is contingent upon the volume and quality of the targeted language data within the ChatGPT training dataset. These systems have well-known downsides, most notably in term of social impact and inefficiency (see Solaiman et al. [6] and references therein), but their value as assistive tools for this task is a good indication of how much systems developed specifically for this task could achieve (while avoiding said downsides).

The novelty of the solution is that all of these necessary resources and goals might exist outside the scope of NLP research. Investment may be driven entirely by internal goals at the community level, or by a linguist for scientific purposes. Moreover, ARBRES was created by a formal linguist, but it doesn't have to be: as long as the grammar is written to teach humans about the language, the required amount of diversity will be found in the data. The resource can then be incrementally built as an educational and/or scientific resource in a form adapted to its audience. On the scale of small language communities, this avoids monopolising experts to create resources that would not be usable by the general public. The more specialised annotations of the data (grammatical categorisation, lemmatisation, coding of consonant mutations) remains inconspicuous, and just serves navigation of the human reader.

The development of wikigrammars is particularly recommended for the construction of pilot project resources on languages with restricted corpora, since even where tech actors fail momentarily to provide finalized tools for speakers, the investment will remain beneficial for the speaking community, which can truly continue to improve the wikigrammar for itself.

Descriptive and formal linguists set themselves the task of producing language analysis material, and can develop these without NLP training. The wiki syntax has a very low entry cost, which is now roughly that of a normal word processing program. In languages with restricted corpora, linguists and trained experts are often very committed to their empirical domain and to the speakers who produce the data. They usually have a precise cultural knowledge, including the diversity of live data, and this also has a beneficial impact on the chosen examples. In terms of human resources, this solution makes it possible to capture their fine-grained expertise. Finally, the wiki solution is designed for large-scale collaboration of potentially isolated contributors. This is particularly suitable for minoritized languages where linguists and trained experts are usually few in numbers, sometimes in precarious socioeconomic situations. Finally, wikigrammars allow for the corpus to be built under the review, direct and indirect, of the entire speaking community.

4 SOCIAL ENGAGEMENT IN MINORITY LANGUAGES

4.1 Public engagement

Internal statistical tools, as well as external analytics systems provide precise insights about the uses of the website, by tracking (anonymously) the more than 100 daily human visits in ARBRES. The graph in Figure 1 shows global visit statistics from October 2023 to the end of January 2024.

Studying the flow of readers makes it possible to identify and understand gaps. One can see the successful entry pages, those that receive the lesser engagement or the shortest reading times, or the particular requests made on search engines that led the readers to the grammar.

Once the website has reached a critical size and a good representation in search engines, the geographical sources of connections can be analysed to provide information on the readership. ARBRES is predominantly utilized within Brittany and among diaspora communities, as seen in Figure 2, which reports the number of visits by city in January 2024.

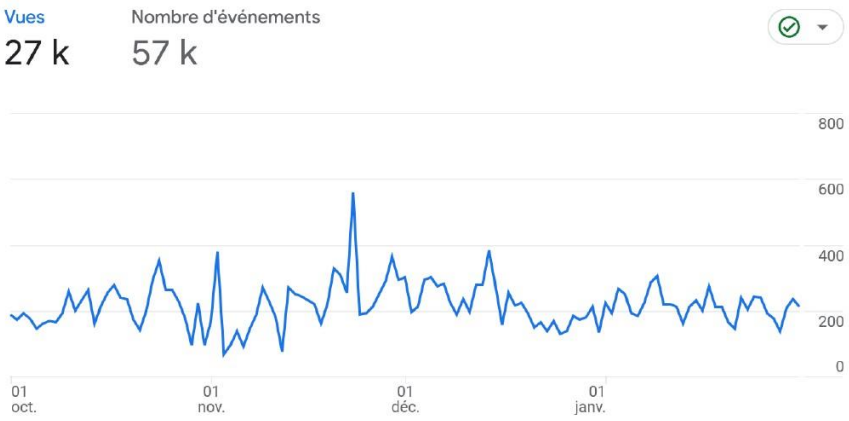


Figure 1: Number of visits on ARBRES from October 2023 to the end of January 2024.

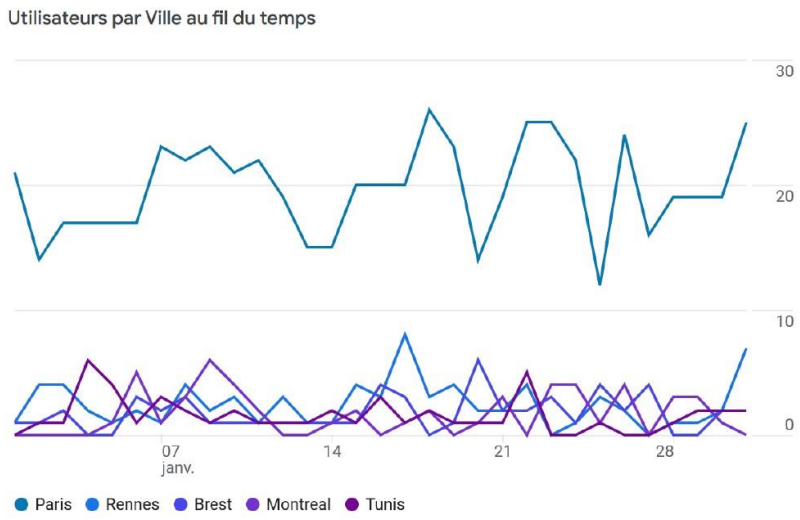


Figure 2: Number of visits on ARBRES in January 2024 split by city

The usage for the website is notably in sync with academic calendars. The sections dedicated to more complex aspects of formal linguistics, offering foundational information in French, experience significant surges in traffic

during typical examination periods in French-speaking regions (e.g., Switzerland, Morocco, Qu´ebec, Algeria, Belgium, etc.).

The precision of the geographic data allows for the observation of the resource's international usage, such as when Breton language courses refer to it. For instance, in 2010, Anna Mouradova started teaching Breton in Moscow, which sparked a spike in connections. Interestingly, one can also see where the resource is not identified (like the sporadic Breton classes in Harvard).

4.2 Equip the interface between science and society

The Breton wikigrammar ARBRES is an experiment of open and participative science (see Jouitteau [7] for an early analysis of the deployment). Wikigrammars bring the scientific process closer to the public. Like any other grammar in open access, it makes available the results of research at the end of its process at a given time. But it also does much more. In synchrony, it links the work to the used sources and to the scientific community. It also sheds light on the past of its making, and on the future of its making. We now illustrate these three dimensions.

Scientific monitoring makes it possible to feed the grammar with the results of the latest research. This effect only derives from its use as a research notebook. The external resources are summarized, referenced and, when open access allows for it, directly linked to. All of these operations bring the readership closer to scientific stakeholders, making them more understandable and more accessible. In 2014, the organization of the Redadeg (Race for the Breton language event) asked for the translation of "I speak Breton, and you?" in different languages. In a few days, linguists from all over the world happily participated in contributing to the page I speak Breton, what about you?, bringing together translations of this sentence in 77 different languages. In support of the event, 1,695 Breton speakers posted self-portraits online with these sentences. The international community of linguists was rendered visible to the general Breton speaking community, and conversely the Breton language appeared very concretely as a production of living speakers to the scientists.

A wikigrammar also includes its whole history. It references the making of its own research. The wiki history function offers for each page a full traceability of the process of knowledge building and data gathering: contributions, corrections, discussions, exploration of new datasets, integration of new bibliographic sources and new hypotheses on the rise being tested. Each page is associated with a complete history giving all modifications made to it since its creation. One can trace back how science is done, how new data and new publications change our hypotheses. The diversity of contributors or lack thereof for each topic is visible. Every contribution is visible and can be duly credited.

Scientific research is the result of a methodology, and is at heart a process accessible to anyone, as long as the methodology is respected. Within these limits the wiki software is designed to allow both cumulative collaboration (massive aggregation of small contributions into a single architecture), and a distributive collaboration (with differentiated tasks). Various competences can then come in together to build a strong resource for the community. This medium raises for the reader the question of their place in the process, enabling a spectrum of engagement levels, ranging from passive activities such as reading, to active participation like commenting, correcting, providing input, writing, linking and so forth. This is particularly welcome in the case of minoritized languages, where speakers commonly report feelings of dispossession of what they consider their language.

Finally, let us discuss a marginal but beneficial effect. Society is rife with sometimes poorly informed debates about languages and especially minoritized languages, due to a lack of verifiable information, a lack of objective knowledge of the linguistic varieties, or an accumulation of inaccuracies. The wikigrammar hosts linguistic

discussion articles which provide concrete elements of analysis on these debates, and proper scientific references. The digital format of these articles makes them directly shareable on social networks, in a format open to a scientific discussion, within the limits of scientific argumentation. In ARBRES, the article about the Sapir-Whorf hypothesis is the second most frequently visited page on the website.

In turn, the traffic generated by this supports search engine optimization and maintains visibility for a work related to marginalized languages on the Internet.

REFERENCES

[1] Mélanie Jouitteau. 2009–2024. ARBRES, Wikigrammaire Des Dialectes Du Breton et Centre de Ressources Pour Son Étude Linguistique Formelle. Retrieved from http://arbres.iker.cnrs.fr

[2] Magnus Manske and Lee Daniel Crocker. 2002. MediaWiki. Retrieved from https ://www.mediawiki.org/wiki/MediaWiki

[3] Mélanie Jouitteau, Yidi Jiang, Yingzi Liu, Salomé Chandora, Kim Gerdes, Bruno Guillaume, Adrien Said-Housseini and Sylvain Kahane. 2022–2024. Autogramm/Breton II. Retrieved from https://github.com/Autogramm/Breton

[4] Mélanie Jouitteau and Reun Bideault. 2023. Outils Numériques et Traitement Automatique Du Breton. In: Langues Régionales de France: Nouvelles Approches, Nouvelles Méthodologies, Revitalisation. Société Linguistique de Paris, 37–74.

[5] Loïc Grobol, and Mélanie Jouitteau. 2024. ARBRES Kenstur: A Breton-French Parallel Corpus Rooted in Field Linguistics. In: Forthcoming.

[6] Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker,Su Lin Blodgett, Hal Daumé III au2, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait and Apostol Vassilev. 2023. Evaluating the Social Impact of Generative AI Systems in Systems and Society. Jun. 12, 2023. arXiv: 2306.05949.

[7] Mélanie Jouitteau. 2012. La linguistique comme science ouverte. In: Lapurdum. Euskal ikerketen aldizkaria | Revue d'études basques | Revista de estudios vascos | Basque studies review 16 (16 1st Oct. 2012), 93–115. doi: 10.4000/lapurdum . 2357.

Welsh Automatic Text Summarisation

JONATHAN MORRIS

Cardiff University

IGNATIUS EZEANI

Lancaster University

IANTO GRUFFYDD

Bangor University

KATHARINE YOUNG

Cardiff University

LYNNE DAVIES

Cardiff University

MAHMOUD EL-HAJ

Lancaster University

DAWN KNIGHT

Cardiff University

Text summarisation is a digital approach to summarising 'key' information contained within texts, and the creation of shortened versions of texts based on this content. Text summarisation function is to provide succinct and coherent summaries to users, something that is often time-consuming and difficult to conduct manually. This is useful in the modern digital world where the creation and sharing of text is ever-increasing, as it enables users to navigate, and make sense of, the dearth of digital information that is available, with ease. This paper reports on work on a project which aims to develop an online Automatic Text Summarisation tool for the Welsh language, ACC (*Adnodd Creu Crynodebau*). This paper contextualises the need for this text summarisation tool, underlines how a dataset for training and testing the methods was created, and outlines plans for the development of the summariser.

Keywords: Text summarisation, Welsh language, Dataset extraction, creation and evaluation

1 INTRODUCTION

Work on automatic text summarisation has a long history in Natural Language Processing (NLP). This work originally focused only on English, as a global lingua franca, but is now used in a range of other language contexts, including French, Spanish, Hindi Arabic, amongst others. The MultiLing project and associated conference series, for example, are a noteworthy champion of developing text summarisation in a range of the world's 7,000+ different languages. The website, http://multiling.iit.demokritos.gr provides an open repository for summarisation tasks test/training data, model summaries, amongst others.

Missing from current summarisation resources are tools that effectively work with the Welsh language. The development of ACC contributes to work on text summarisation in minority languages and contributes to the technological resources available to Welsh speakers.

2 WELSH LANGUAGE ONLINE

There exists a relatively low use of Welsh language websites and e-services, despite the fact that numerous surveys suggest that Welsh speakers would like more opportunities to use the language, and that there has been an expansive history of civil disobedience in order to gain language rights in the Welsh language context [1].

One reason for the relatively low take-up of Welsh-language options on websites is the assumption that the language will be too complicated [1]. Concerns around the complexity of public-facing Welsh language services and documents are not new. A series of guidelines on creating easy-to-read documents in Welsh are outlined in *Cymraeg Clir* [2]. Williams [2] notes that the need for simplified versions of Welsh is arguably greater than for English considering:

1. many Welsh public-facing documents are translated from English,
2. the standard varieties of Welsh are further removed from local dialects compared to English, and
3. newly-translated technical terms are more likely to be familiar to the reader.

The principles outlined in Cymraeg Clir therefore include the use of shorter sentences, everyday words rather than specialised terminology, and a neutral (rather than formal) register [2].

ACC will provide the means for summarising and simplifying digital language sources which will help to address the fears of Welsh speakers that language online is too complicated.

ACC will also contribute to the digital infrastructure of the Welsh language. The most recent Welsh Government strategy for the revitalisation of Welsh has infrastructure (and particularly digital infrastructure) as a main theme (along with increasing the number of speakers and increasing language use [3]). The aim is to "ensure that the Welsh language is at the heart of innovation in digital technology to enable the use of Welsh in all digital contexts" [3]. Given the introduction of Welsh Language Standards (see [4]) and a concerted effort to both invest in Welsh language technologies and improve the way in which language choice is presented to the public, the development of ACC will complement the suite of Welsh language technologies (e.g. [5]) for both content creators and Welsh readers.

It is also envisaged that ACC will contribute to Welsh-medium education by allowing educators to create summaries for use in the classroom as pedagogical tools. Summaries will also be of use to Welsh learners who will be able to focus on understanding the key information within a text.

3 DATA EXTRACTION

The first stage of the development process is to develop a small corpus (dataset) of target language data that will subsequently be summarised and evaluated by human annotators and used to develop and train the automated summarisation models (i.e. acting as a gold-standard dataset).

Wikipedia[1] was selected as the primary source of data for creating the Welsh language dataset for ACC. This was owing to the fact that an extensive number of Welsh language texts exist on this website (over 133,000 articles), all of which are available under GNU Free Documentation license. To ensure that pages that contained a sufficient quantity of text were extracted for use, a minimum threshold of 500 tokens per article and a target of at least 500 articles was established at the outset. A selection of 800 most accessed Wikipedia pages in Welsh were initially extracted for use. An additional 100 Wikipedia pages were included from the WiciAddysg[2] project organised by the

[1] Welsh Wikipedia: https://cy.wikipedia.org/wiki/Hafan (Wicipedia)
[2] WikiAddysg: https://cy.wikipedia.org/wiki/Categori:Prosiect_WiciAddysg

National Library of Wales and Menter Iaith Môn . However, it was observed that more than 50% of the articles from this original list of Wikipedia pages did not meet the minimum-token threshold of 500. To mitigate this, a list of 20 Welsh keywords was used to generate an additional 100 Wikipedia pages per keyword (which was provided by the first author and contained words synonymous with the Welsh language, Welsh history and geography). This was added to the list of 100 most-edited Welsh Wikipedia pages and pages from the WiciAddysg project.

The data extraction applied a simple iterative process and implemented a Python script based on the WikipediaAPI[3] that takes a Wikipedia page; extracts key contents (article text, summary, category) and checks whether the article text contains a minimum number of tokens. At the end of this process, the dataset was created from a total of 513 Wikipedia pages that met the set criteria. The extracted dataset contains a file for each Wikipedia page with the following structure and tags:

```
<title>Article Title on Wikipedia</title>

<text> Article Text </text>

<category>Article Categories </category>.
```

These files are available in both plain text and html file formats.

4 DATASET CREATION

A total of 19 undergraduate and postgraduate students from Cardiff University were recruited to create, summarise and evaluate the dataset. Of these students, 13 were undertaking an undergraduate or postgraduate degree in Welsh which involved previous training on creating summaries from complex texts. The remaining six students were undergraduate students on other degree programmes in the Humanities and Social Sciences at Cardiff University and had completed their compulsory education at Welsh-medium or bilingual schools.

Students were asked to complete a questionnaire prior to starting work which elicited biographical information. A total of 17 students had acquired Welsh in the home. One student acquired the language via Welsh-medium immersion education and one student had learned the language as an adult. The majority of students came from south-west Wales (n=11). This region included the counties of Carmarthenshire, Ceredigion, Neath Port Talbot, and Swansea. A further five students came from north-west Wales which comprised the counties of Anglesey and Gwynedd. One student came from south-east Wales (Cardiff), one from mid Wales (Powys), and one from north-east Wales (Conwy).

A broad distinction can be made between northern and southern Welsh. The two varieties (within which further dialectal differences exist) exhibit some differences at all levels of language structure although all varieties are mutually intelligible. Students were asked four questions which elicited information on the lexical, grammatical, and phonological variants they would ordinarily use. The results largely corresponded to geographical area: 11 students used southern forms and seven students used northern forms (including the student from mid Wales). One student, from Cardiff, used a mixture of both northern and southern forms.

Students were given oral and written instructions on how to complete the task. Specifically, they were told that the aim of the task was to produce a simple summary for each of the Wikipedia articles (allocated to them) which contained the most important information. They were also asked to conform to the following principles:

- The length of each summary should be 230 - 250 words.

[3] Wikipedia API: https://pypi.org/project/wikipedia/

- The summary should be written in the author's own words and not be extracted (copy-pasted) from the Wikipedia article.
- No information which is not included in the article should be included in the summary.
- Any reference to a living person in the article should be anonymised in the summary (to conform to the ethical requirements of each partner institution).
- All summaries should be proofread and checked using spell checker software (Cysill[4]) prior to submission.

Further instruction was given on the register to be used in the creation of summaries. Students were asked to broadly conform to the principles of Cymraeg Clir [2] and, in particular, avoid less common short forms of verbs and the passive mode, and use simple vocabulary where possible instead of specialised terms.

Each student completed between 60 - 100 summaries between July and October 2021. The median amount of time spent on each summary was 30 minutes. The complete dataset comprises 1,461 summaries with the remaining 39 summaries not being completed due to one student prematurely dropping out of the project and some instances of unsuitable articles (e.g. lists of bullet points).

Three of the postgraduate students recruited were also asked to evaluate the summaries by giving a score between one and five. Table 1 shows the marking criteria.

Table 1: Criteria for the marking of summaries

Score	Criteria
5	Very clear expression and very readable style. Very few language errors. Relevant knowledge and a good understanding of the article; without significant gaps.
4	Clear expression and legible style. Small number of language errors. Relevant knowledge and a good understanding of the article, with some gaps.
3	Generally clear expression, and legible style. Number of language errors. The knowledge and understanding of the article is sufficient, although there are several omissions and several errors.
2	Expression is generally clear but sometimes unclear. Significant number of language errors. The knowledge and understanding of the article is sufficient for an elementary summary, but there are a number of omissions and errors.
1	Expression is often difficult to understand. Defective style. Persistently serious language errors. The information is inadequate for summary purposes. Obvious deficiencies in understanding the article.

Both the mean and median scores for the summaries were 4. Evaluators were instructed to fix common language errors (such as mutation errors and spelling mistakes) but not to correct syntax.

[4] Cysill: https://www.cysgliad.com/cy/cysill/

5 SUMMARISATION TOOL DESCRIPTION

The second phase of this summarisation project is to use the corpus dataset to inform the iterative development and evaluation of digital summarisation tools. The main approaches to text summarisation include extraction-based summarisation and abstraction-based summarisation. The former extracts specific words/phrases from the text in the creation of the summary, while the latter works to provide paraphrased summaries (i.e. not directly extracted) from the source text. The successful extraction/abstraction of content, when using summarisation tools/approaches, depends on the accuracy of automatic algorithms (which require training using hand-coded gold-standard datasets).

As an under-resourced language with limited literature on Welsh summarisation, applying summarisation techniques from the literature helps in having initial results that can be used to benchmark the performance of other summarisers on the Welsh language. In this project we are to develop a combination of extractive and abstractive single-document summarisation methods. The process will start by implementing and evaluating basic baseline systems that are frequently used in the literature as bench-lines. These will be followed by more complex state of the art summarisation models as well as hybrid systems as toplines.

5.1 Baselines

The sections below provide an overview of the summarisation systems that this project will be focusing on currently as well as within the life of the project.

5.1.1 First Sentence Summariser

Rather than using a document's title or keywords, some summarisers tend to use the first sentence of an article to identify the topic to be summarised. The justification behind selecting the first sentence as being representative of the relevant topic is based on the belief that in many cases, especially in news articles or articles found on Wikipedia, the first sentence tends to contain key information about the content of the entire article [6, 7, 8].

5.1.2 TextRank

This summarisation technique was introduced by Rada Mihalcea and Paul Tarau [9]. This was the first graph-based automated text summarisation algorithm that is based on the simple application of the PageRank algorithm. PageRank is used by Google Search to rank web pages in their search engine results [10]. TextRank utilises this feature to identify the most important sentences in an article.

5.1.3 LexRank

Similar to TextRank, LexRank uses a graph-based algorithm for automated text summarisation [11]. The technique is based on the fact that a cluster of documents can be viewed as a network of sentences that are related to each other. Some sentences are more similar to each other while some others may share only a little information with the rest of the sentences. Like TextRank, LexRank too uses the PageRank algorithm for extracting top keywords. The key difference between the two baselines is the weighting function used for assigning weights to the edges of the graph. While TextRank simply assumes all weights to be unit weights and computes ranks like a typical PageRank execution, LexRank uses degrees of similarity between words and phrases and computes the centrality of the sentences to assign weights [11].

5.2 Toplines

As the project progresses, we will develop more complex summarisers and evaluate their performance by comparing the summarisation results of the three baselines mentioned above. The purpose of the topline summarisers is to prove that using language related technology to summarise Welsh documents will improve the results of those produced by the baseline summarisers.

5.2.1 TF.IDF Welsh Summariser

A summariser using Text Frequency Inverse Document Frequency (TF.IDF) works on findings words that have the highest ratio of those words frequency in the to be summarised document in comparison to their occurrence in the full set of documents to be summarised [12]. TF.IDF is a simple numerical statistic which reflect the importance of a word to a document in a text collection or corpus and is usually used as a weighing factor in information retrieval, thus using it to find important sentences in extractive summarisation [13, 14].

The summariser will work on finding key and important words in the documents to be summarised in an attempt to produce relevant summaries. Using TF.IDF in the Welsh language is not new. Arthur et al. [15], used a social network that they built using Twitter's geo-locations to identify contiguous geographical regions and identify patterns of communication within and between them. Similarly, we will use TF.IDF to identify important sentences based on patterns detected between the summarised document and the summaries corpus.

5.2.2 TF.IDF Welsh Summariser with Welsh Word Embeddings

In order to improve the similarity measure between sentences, we use pre-trained word embedding features combined with the previously mentioned TF.IDF features. For that we use the FastText Welsh pre-trained word vectors [16]. FastText is an extension of the word2vec [17] model where instead of learning vectors for words directly, FastText represents each word as an n-gram of characters, which helps in capturing the meaning for shorter words and allow the embeddings to understand suffixes and prefixes. Ezeani et al. [18] leveraged existing language models such as Welsh FastText for multi-task classification of Welsh part of speech and semantic tagging. We will repeat the experiment but this time using Welsh word embeddings created by Corcoran et al. [19] where word2vec and FastText is used to automatically learn Welsh word embeddings taking into account syntactic and morphological idiosyncrasies of this language. We will build upon those two previous efforts and harness the language models towards enriching the performance of the TF.IDF summariser in 5.2.1.

5.3 State of the Art Welsh Summarisers

The final stage of the project is to use state of the art summarisation technologies to summarise Welsh documents. This will include building Extractive and Abstractive summarisers using deep neural network machine learning techniques or what is known as Deep Learning. The summarisation state of the art literature shows a great shift towards using deep learning to create extractive and abstractive supervised and unsupervised summarisers using deep learning models such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short Term Memory (LSTM) and many others [20, 21, 22 ,23]. In this project we will combine the use of the aforementioned Welsh word embeddings to try and improve the results and create Welsh summarisation systems that are on par with other English and European state of the art summarisers.

6 EVALUATION

The gold-standard summaries created by the human summarisers as described in Section 4 will be used to automatically evaluate any system summaries generated by the models developed in Section 5. The system summaries will be evaluated using the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics [24]. ROUGE is a set of metrics used for evaluating automatic summarisation in natural language processing. The metrics compare an automatically produced summary against gold-standard summaries.

As an additional level of evaluation, a sample of system summaries generated by our best performing summarisation model (see Section 5.2) will be manually evaluated by native Welsh speakers in order to measure the quality of those summaries. The final stages of this project will include the development of a freely-available user-friendly web-based user interface that can be used by users from all age groups. The system will allow users to define the level of compression (e.g. a summary of no more than 200 words). The summariser will also be available as open-source Python packages to allow developers to work on enhancing the summarisers in the future.

7 CONCLUSION

The released version of ACC will contribute to the automated tools available in the Welsh language and facilitate the work of those involved in document preparation, proof-reading, and (in certain circumstances) translation. The tool will also allow professionals to quickly summarise long documents for efficient presentation. For instance, the tool will allow educators to adapt long documents for use in the classroom. It is also envisaged that the tool will benefit the wider public, who may prefer to read a summary of complex information presented on the internet or who may have difficulties reading translated versions of information on websites. To keep up to date with developments on this tool, please visit the main project website at: https://corcencc.org/acc/

ACKNOWLEDGMENTS

This research was funded by the Welsh Government, under the Grant “Welsh Automatic Text Summarisation”. We are grateful to Jason Evans, National Wikimedian at the National Library of Wales, for his initial advice.

REFERENCES

[1] Jeremy Evas and Daniel Cunliffe. 2016. Behavioural Economics and Minority Language e-Services—The Case of Welsh. In Durham, M. and Morris, J. (eds), Sociolinguistics in Wales. Palgrave Macmillan. London. 61-91.

[2] Cen Williams. 1999. Cymraeg Clir: Canllawiau Iaith. Bangor: Gwynedd Council, Welsh Language Board and Canolfan Bedwyr.

[3] Welsh Government. 2017. Cymraeg 2050 - A million Welsh speakers. Retrieved from https://gov.wales/sites/default/files/publications/2018-12/cymraeg-2050-welsh-language-strategy.pdf

[4] Patrick Carlin and Diarmait Mac Giolla Chríost. 2016. A standard for language? Policy, territory, and constitutionality in a devolving Wales. In Durham, M. and Morris, J. (eds), Sociolinguistics in Wales. Palgrave Macmillan. London. 93-119.

[5] Uned Technolegau Iaith Prifysgol Bangor. 2021. Cysgliad: Help i ysgrifennu yn Gymraeg. Retrieved from https://www.cysgliad.com/cy/

[6] Dragomir R. Radev, Hongyan Jing, Małgorzata Styś and Daniel Tam. 2004. Centroid-based Summarization of Multiple Documents. Information Processing and Management, 40, 919–938.

[7] M. Fattah and F. Ren. 2008. Automatic Text Summarization. In Proceedings of World Academy of Science, 27, World Academy of Science, 192–195.

[8] Jen-Yuan Yeh, Hao-Ren Ke and Wei-Pang Yang. 2008. iSpreadRank: Ranking sentences for extraction-based summarization using feature weight propagation in the sentence similarity network. Expert Systems with Applications, 35(3), 1,451–1,462.

[9] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing Order into Text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 404-411.

[10] Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems, 30(1-7), 107-17.

[11] Erkan, G. and Radev, D. R. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence

research, 22, 457–479.

[12] Gerard Salton and Michael J. McGill. 1986. Introduction to modern information retrieval. McGraw Hill. New York.

[13] Hajime Mochizuki and Manabu Okumura. 2000. A Comparison of Summarization Methods based on Taskbased Evaluation. In Proceedings of the 2nd International Conference on Language Resources and Evaluation. Athens, Greece, 404-411.

[14] C. G. Wolf, S. R. Alpert, J. G. Vergo, L. Kozakov and Y. Doganata. 2004. Summarizing Technical Support Documents for Search: Expert and User Studies. IBM Systems Journal, 43(3), 564–586.

[15] Rudy Arthur and Hywel T. P. Williams. 2019. The human geography of Twitter: Quantifying regional identity and inter-region communication in England and Wales. PloS one, 14 (4), e0214466.

[16] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. Dec. 12, 2016. arXiv:1612.03651.

[17] Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. Jan. 16, 2013. arXiv:1301.3781.

[18] Ignatius Ezeani, Scott Piao, Steven Neale, Paul Rayson and Dawn Knight. 2019. Leveraging pre-trained embeddings for Welsh taggers. In Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). Florence, Italy, 270-280.

[19] Padraig Corcoran, Geraint Palmer, Laura Arman, Dawn Knight, and Irena Spasić. 2021. Creating Welsh language word embeddings. Applied Sciences, 11(15), 6896.

[20] Shengli Song, Haitao Huang and Tongxiao Ruan. 2019. Abstractive text summarization using LSTM-CNN based deep learning. Multimedia Tools and Applications, 78(1), 857-875.

[21] Nadhem Zmandar, Abhishek Singh, Mahmoud El-Haj and Paul Rayson. 2021. Joint abstractive and extractive method for long financial document summarization. In Proceedings of the 3rd Financial Narrative Processing Workshop. Lancaster University, Lancaster, England, 99-105.

[22] Mahmoud El-Haj, Nadhem Zmandar, Paul Rayson, Ahmed AbuRa'ed, Marina Litvak, Nikiforos Pittaras, George Giannakopoulos, Aris Kosmopoulos, Blanca Carbajo-Coronado and Antonio Moreno-Sandoval. 2021. The Financial Narrative Summarisation Shared Task FNS 2021. In Proceedings of the 3rd Financial Narrative Processing Workshop. Lancaster University, Lancaster, England, 120-125.

[23] P. G. Magdum and Sheetal Rathi. 2021. A Survey on Deep Learning-Based Automatic Text Summarization Models. Advances in Artificial Intelligence and Data Engineering. Springer. Singapore. 377-392.

[24] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out. Association for Computational Linguistics. Barcelona, Spain, 74–81.

What is Required to get Workable ASR for Cornish

PREBEN VANGBERG

Bangor University, Wales

LEENA SARAH FARHAT

Bangor University, Wales

This research investigates the challenges and potential solutions for constructing a functional automatic speech recognition (ASR) system for the Cornish language. Cornish is a Brittonic language formerly spoken in Cornwall, England, but declined in popularity in the 18th century. Cornish has had a resurgence in recent decades, and there is currently a rising need for an ASR system that can recognise and transcribe Cornish speech. This research cites two major problems in constructing an ASR system for Cornish: a dearth of training data and the language's unusual phonology. This preliminary study provides a variety of potential solutions to these issues, including employing zero-shot learning, training a model on a modified Cornish orthography, and using phonemes as an intermediary stage in the transcription process. This paper also explores the necessity for a publicly accessible transcribed audio dataset for Cornish and suggests that Mozilla Common Voice could be an appropriate platform for hosting such a dataset. The research closes by proposing that a mix of the methodologies offered be utilised to create a functional ASR system for Cornish but recognises the need for more work in this area.

Keywords: Automatic Speech Recognition (ASR), Cornish language, Cornish orthography, Speech Technology

1 INTRODUCTION

This paper introduces the challenges and potential approaches for developing a workable Automatic Speech Recognition (ASR) system for the Cornish language with no appropriate data for the task.

1.1 The Cornish language

Cornish is a Brittonic language that is spoken in Cornwall. It is connected to Breton and Welsh and is less related to Scottish Gaelic, Manx, and Irish Gaelic. Cornish was formerly the most widely spoken language in Cornwall, but it fell out of favour following the English invasion in the 13th century. While the Cornish language endured a dramatic collapse in the 18th century, it has seen a remarkable recovery in recent decades, thanks to enthusiastic individuals and organisations dedicated to preserving and promoting Cornish culture. Cornish revitalization has been fraught with difficulties, among them the lack of a recognized standard for writing the language. Several different orthographies were proposed, but this was not resolved until 2008, when there was general agreement in favour of the Standard Written Form (SWF), but with the other popular variations such as Kernewek Kemmyn, Unified Cornish, Tudor Cornish and Revived Late Cornish also recognized [1]. 563 people noted that they speak Cornish in the 2021 UK census [2], while Cornish learners number in the low thousands. This small pool of speakers makes it difficult to collect enough speech data to adequately train ASR models to develop an ASR system for the language. Despite these obstacles, the Cornish language community has been active in creating content for their language. Initiatives such as Cornish Language Radio and Cornish Language TV allow Cornish people to utilise the language in everyday situations, enhancing the linguistic environment. While not directly adding to the pool of native speakers, the considerable number of Cornish learners plays an important part in the language's resuscitation.

Surprisingly, the flood of learners can also alter Cornish pronunciation. Because students come from a variety of linguistic origins, they may bring new pronunciations or variants on existing sounds. This tendency, termed as

'interference,' is a normal element of language interaction and can lead to a language's growth over time. The small number of native speakers in Cornish can mean that the effect of learners may be more obvious. This effect, however, is not always harmful; it can contribute to the language's plasticity and durability.

The Cornish language community is continually working to fix pronunciation difficulties and ensure uniformity in the language's norms. Overall, the revival of Cornish is a remarkable monument to the lasting power of language and the determination of individuals and communities devoted to preserve their linguistic heritage. While obstacles exist, notably in assembling a large speech corpus for ASR systems, the Cornish language community is actively tackling these concerns and making great progress in revitalising this unique and precious language.

1.2 Automatic Speech Recognition

ASR models must be trained on huge datasets of labelled speech. Collecting enough labelled data to train an efficient ASR model in languages with minimal resources, such as Cornish, can be challenging, if not impossible. There has been a surge of interest in building ASR systems for under-resourced languages in recent years. This is partly due to the development of deep learning, which has enabled ASR models to be trained with less data. Furthermore, there is a rising recognition of the need for language preservation, and ASR can play a role in preserving under-resourced languages. This paper seeks to investigate what is required to enable workable ASR for the Cornish language.

2 TRAINING DATA FOR ASR AND OTHER SPEECH TECHNOLOGIES

The creation of an ASR system normally requires two key components, an acoustic model responsible for converting speech to text, and a language model responsible for fixing errors in this text. These are trained in separate ways and require different data.

2.1 Training data for acoustic models

The occurrence of many variants of Cornish, especially ancient versions, can affect pronunciation. Speakers may adhere to the pronunciation standards of the variety with which they are familiar, resulting in differences in how specific words or phrases are uttered. This is a normal phenomenon in any language with many variations, and it does not necessarily jeopardise the language's unity. When training ASR models, it is critical to expose the models to a wide range of Cornish pronunciations, particularly those linked with various orthographic variations. This exposure will aid the models' generalisation and performance when confronted with speakers who employ non-standard pronunciations. The only data that is quite specific for the ASR application is transcribed speech. This is used to train the acoustic part of the overall ASR system. While speech synthesis also uses transcribed speech, there is often a difference in quality. Speech synthesis requires high-quality audio without noise so that the voice becomes clearer. ASR on the other hand can often benefit from noisy audio as it makes the models more robust. There is no publicly available transcribed audio dataset available for Cornish. Academics do have access to recordings of readings and interviews through the medium of Cornish (see work undertaken by Szczepankiewicz [3]). However, while some of this data might be transcribed, it has not yet been segmented into sentence-length clips usable for training acoustic models. There exist other transcribed recordings of speakers using the language, however, most of these are licensed and not available for use in ASR.

2.1.1 Mozilla Common Voice

Mozilla Common Voice [4] is a crowd-sourced speech corpus developed and maintained by Mozilla. It has datasets for 114 languages and the number of datasets is increasing. Cornish does not currently have an active Common Voice dataset. There are efforts to try to get a Cornish dataset up and going. The first step to launching a dataset on Common Voice is to localise the Common Voice interface into the target language. As of 30 October 2023, 56% of the interface has been localised. In addition to this, Common Voice requires a set of Creative Commons Zero (CC0) licensed sentences for users to record. Getting a Cornish Common Voice dataset on Common Voice represents a significant opportunity for the language and Cornish ASR as it would encourage speakers to contribute data for the development of language technologies in an easy-to-use and crowd-sourced manner.

2.2 Training data for language models

Language models mainly use plain text as training data. Fortunately, there is a substantial amount of Cornish text on the internet that can be used for this purpose, but it is not easy to find or use. Often, we had to rely on the Internet Archive to get enough Cornish text to create usable language models. Wikipedia has a significant amount of Cornish text that could be used, however, it requires extensive cleaning due to the nature of the data. This is not impossible and has been done for other languages (e.g. it has been done for Breton in the past [5]), but it is something that has yet to be done for Cornish.

2.2.1 Labelling Cornish text

As noted above, Cornish orthography has recently been standardised into a single standard form, which benefits the language's uniformity and accessibility. It is crucial to note, however, that the earlier variants of Cornish spelling that still exist are used by certain speakers and learners. As a result, striking a balance between sticking to the SWF standard and admitting sentences in other variations is critical. This inclusive approach will help capture the language's depth and vibrancy while supporting the acceptance of the SWF standard. Something that requires more research is the impact of the various competing orthographies for Cornish on the effectiveness of language models. It might be that having separate language models targeting the various orthographies yields better results than one combined model. To be able to investigate this, Cornish datasets need to include information about which orthography is used. With enough classified text, it might be possible to automate this classification to a certain extent using tools such as Tawa [6].

3 METHODOLOGY

Several methodologies were trialled for this project. These methods were selected in accordance with the linguistic background of Cornish as well as considering the data scarcity.

3.1 Zero-shot

Zero-shot learning allows models to execute tasks without being explicitly trained on them. Unlike standard supervised learning, which trains models on a collection of labelled examples, zero-shot learning models may generalise to new tasks and data with no extra training. Previous experiments have shown that using acoustic models designed for closely related languages and dialects paired with bespoke language models can produce decent results [7]. Therefore some preliminary tests were run using the Wav2Vec2 framework [8] and using models trained for various languages. These models can be seen in Table 2. In addition to running tests using various

acoustic models, different configurations of text were tested to train our language models. The texts used come from Cornwall Council (referred to as Corpus Kernewek in the rest of the paper) and a dump of the Cornish Wikipedia (See https://dumps.wikimedia.org/kwwiki/latest/). Each one was tested on its own in addition to a combined model.

3.2 Training using a modified orthography

The closest related language to Welsh is Cornish. A substantial amount of words and phrases derive from their common Brittonic ancestor, and some regular orthographical changes enable the relationship between them to be traced. For example, in Welsh, the sounds /v/ and /f/ are written as ⟨f⟩ and ⟨ff⟩ respectively, while in Cornish these are written as ⟨v⟩ and ⟨f⟩ like in English. Cornish has a complicated phonology, with some characteristics that set it apart from other Brittonic languages. The assimilation of Old Cornish ⟨d⟩ to Middle Cornish ⟨s⟩, for example, Old Cornish *dad* or *father* and Middle Cornish *tas*, is one of the most distinguishing aspects of Cornish. This shift is believed to have occurred about the 11th century [9].

Another characteristic of Cornish phonology is the palatalization of /d/ to /dʒ/ before front vowels, as in Middle Cornish *dzadn* (tooth) and Modern Cornish /dʒayn/. This shift is believed to have occurred in the 13th century, approximately. Cornish also features a variety of distinct vowel phonemes, including the short vowels /I/ and /oe/ and the long vowels /I:/ and /oe:/. These vowels are supposed to have evolved from the diphthongization of Old Cornish i and u preceding nasal consonants.

Due to the zero-shot Welsh ASR model making quite recognisable orthographic mistakes, a simple Python function was created to convert common orthographic differences in Welsh to Cornish SWF. The Wav2Vec2 models were then fine-tuned using this 'Cornishified' Welsh data in an attempt to aid the Welsh ASR models in correctly transcribing these phonemes.

3.3 Using phonemes as an intermediate step

The International Phonetic Alphabet (IPA) [10] is a language-independent way of transcribing speech. Another test used was to use IPA as an intermediate step. Since IPA is language-independent, this means that the data could be transcribed to IPA using models trained for that purpose for other languages, and then that IPA translated into Cornish orthography.

3.3.1 Speech to phoneme

The first step is to convert the audio into IPA. There are various ways of doing this. Allosaurus [11] is a piece of software designed for this purpose. Various models can be found on HuggingFace (see https://huggingface.co/). The Wav2Vec2 model by Phy [12] was used as a proof of concept.

3.3.2 Phoneme to grapheme

The next step is a bit more complex. This step is not too dissimilar to a normal translation task; however, it is character-based as opposed to word-based. There were two ways to prototype the systems quickly; the first was a rule-based system and the other was to train a sequence-to-sequence model.

Rule-based approach: The strength of the rule-based approach is that it is quite easy to throw something together and get decent results. The downside, however, is that it is not very 'clever' and it is quite easy to run into problems since it is difficult to take into consideration the context in which a phoneme occurs.

Unfortunately the IPA from the model contains no information about the length of the phonemes, and it was difficult to take into account germination and vowel length. Certain vowels are ambiguous as well. For example, depending on the orthography, /E/ can be written as ⟨e⟩ or ⟨eu⟩, and /i:/ can be written as ⟨i⟩ or ⟨y⟩.

To overcome this the transformation functionality of the Tawa Toolkit [6] was chosen to fix some of these issues by creating a set of transformation rules to convert single consonants into double consonants and resolving ambiguous vowels. The language model used for this was trained from the text from Cornwall Council. These rules were not tested or optimised but decided ad-hoc based on what was estimated to improve the results.

Using sequence-to-sequence models: Parts of the pronunciation dictionary in Akademi Kernewek's online dictionary (see https://www.cornishdictionary.org.uk/) were accessible. It was hypothesised that this could be used to create a sequence-to-sequence model. Several types of LSTMs, Transformers, and other models were tested. However, there was not enough data to train the models effectively, so this approach was abandoned.

4 RESULTS

Overall, the results were promising for this exploratory work. These experiments show that ASR for Cornish is achievable but is still far from optimal.

4.1 Test data used

For testing purposes, only a single Cornish narration of about 15 minutes was accessible. This recording was manually split into segments for testing. Due to privacy concerns, further details about the recording cannot be provided. This test set was challenging because the narration was not simply a reading of text, but rather a dramatized performance. This led to overdramatized pronunciation of some sentences.

4.2 Language Models

The first objective was to investigate the efficiency of the various configurations of our language models. All the language models were optimised and tested using the test data and the techiaith/wav2vec2-xlsr-ft-en-cy model. In total three were tested as shown in Table 1. While the Corpus Kernewek model performed better than the others, it is likely not statistically significant, due to the small number of test cases in the dataset.

Table 1: Overview of the various error rates of the three language models in percent

Text Corpus	WER[1]	WER with LM	CER[2] with LM
Corpus Kernewek	95.05	78.37	43.54
Wikipedia	95.05	83.26	44.93
All	95.05	83.42	45.24

4.3 Zero-shot ASR

For Cornish, one major disadvantage of zero-shot learning is that it might be sensitive to orthography difficulties. Cornish has a complicated orthography with several distinctive peculiarities, including the use of diacritics and the assimilation of *d* to *s*. Many zero-shot learning models have not been taught to deal with spelling concerns, which

[1] Word Error Rate
[2] Character Error Rate

can lead to mistakes. Welsh, as one of the closest related languages to Cornish, has a more standardised spelling than Cornish. However, it differs significantly from that of Cornish. As seen in Table 2, the various languages did not perform significantly differently. However, the results do differ when a language model is included, and techiaith/wav2vec2-xlsr-ft-en-cy outperforms the rest with a decent margin.

4.4 Training using a modified orthography

A range of Wav2Vec2 models were trained using a range of different hyper-parameters. However, for all attempts, training failed to learn anything of value. The final results were a WER of 95.58% and a CER of 86.67%, when using a language model. Note that the reason it is below 100% is that it mostly outputs white space, which means that it sometimes is correct. The reason it failed to learn anything might be due to the data being broken so the results are inconclusive. In conclusion, while there might be a scenario where a better model could be trained, it is not possible to say whether this approach is viable now, and it is a significant likelihood that it is not.

Table 2: Overview of the various error rates for the zero-shot models in percent, sorted by CER.

Language	Model	WER without LM	WER	CER
CY	techiaith/wav2vec2-xlsr-ft-cy	96.47	86.41	59.26
XX	voidful/wav2vec2-xlsr-multilingual-56	95.82	84.46	55.90
BR	DrishtiSharma/wav2vec2-large-xls-r-300m-br-d2	96.90	91.74	53.01
EN	jonatasgrosman/wav2vec2-large-xlsr-53-english	96.90	82.83	51.95
EN-CY	techiaith/wav2vec2-xlsr-ft-en-cy	95.05	78.37	43.54

4.5 Using phonemes as an intermediate step

Our testing found that this system had a WER of 100% and a CER of 48.63% without any language models using an amazingly simple rule-based substitution approach. Given that none of the steps in this process has been optimised, given the lack of support for double consonants, and given the lack of a language model this is quite promising. Testing also showed that including Tawa transformation rules improved the results, but definitive numbers on the improvement are not yet available.

5 CONCLUSIONS AND FUTURE WORK

It is clear that while ASR for Cornish is achievable at this time, it is far from optimal and a significant amount of work needs to be undertaken to improve upon this.

Speech data is essential for Cornish ASR work for several reasons. It allows ASR systems to learn the acoustic properties of Cornish, including the various sounds and how they are combined. Speech data also helps ASR systems learn the statistical relationships between sounds and words, which is necessary for accurate transcription. In addition to training ASR systems, speech data can be used to develop other language technologies, such as text-to-speech and machine translation systems. Increasing the amount of available speech data for Cornish will support the development of new and improved language technologies. Having a Cornish Common Voice dataset on Common Voice would be a significant step forward for the Cornish language and Cornish ASR. It would make it easy for speakers to contribute data for the development of language technologies.

Future work, once there is a dataset that can be used as a foundation for language technologies, includes developing ASR models that are more robust to language variation. Cornish has several variations, so it is important to develop ASR models that can reliably transcribe all of them. One approach would be to use training corpora that would contain data from a number of different variations.

REFERENCES

[1] Albert Bock and Benjamin Bruc. 2008. An Outline of the Standard Written Form of Cornish. Retrieved from https://kernowek.net/Specification_Final_Version.pdf.

[2] Office for National Statistics. 2022. Language, England and Wales - office for national statistics. Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/language/bulletins/languageenglandandwales/census2021

[3] P. Szczepankiewicz. 2023. Linguistic variation in revived Cornish. Research grant 2022/45/N/HS2/00869. National Science Centre, Adam Mickiewicz University. Retrieved from https://projekty.ncn.gov.pl/index.php?projekt_id=557852

[4] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. Mar. 5, 2020. arXiv:1912.06670.

[5] Preben Vangberg and Leena Farhat. 2023. Speech-to-text for Breton. In Celtic student conference, 2023.

[6] William John Teahan. 2018. A Compression-Based Toolkit for Modelling and Processing Natural Language Text. Information 9(12), 294. https://doi.org/10.3390/info9120294

[7] Preben Vangberg and Leena Farhat. 2023. Devisa: Exploring transfer learning in an interdialectal setting for Romansch. Presented at the XIX International Conference on Minority Languages, 2023.

[8] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed and Michael Auli. 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. Oct. 22, 2020. arXiv: 2006.11477.

[9] Talat Chaudhri. 2007. Studies in the Consonantal System of Cornish. PhD Thesis, Department of Welsh and Celtic Studies, Aberystwyth University.

[10] I. P. Association. 1999. Handbook of the international phonetic association: A guide to the use of the international phonetic alphabet. Cambridge University Press. Cambridge.

[11] Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, Alan W Black and Florian Metze . 2020. Universal phone recognition with a multilingual allophone system. In IEEE international conference on acoustics, speech and signal processing (ICASSP) 2020. IEEE, 2020, Barcelona, Spain, 8,249–8,253.

[12] Vitou Phy. 2022. Automatic phoneme recognition on TIMIT dataset with wav2vec 2.0. Retrieved from https://huggingface.co/vitouphy/Wav2Vec2-xls-r-300m-timit-phoneme

How to Work with Lesser Resourced Languages and Large Language Models

COLIN JARVIS

OpenAI

This chapter explores how large language models (LLMs), such as ChatGPT, handle languages with limited training data. A description of how LLMs are trained and how they can be optimized is provided. Using Icelandic error correction and Welsh Text-To-Speech (TTS) as case studies, this chapter outlines the constraints inherent in these models, strategies for refining them, and practical methods for implementing these strategies in languages with limited resources in real-world scenarios.

Keywords: Large Language Models, Optimization, Training Data, Icelandic, Welsh, Lesser Resourced Languages

1 INTRODUCTION

Large language models (LLMs) have excited the imaginations of many since they've entered the public eye with the launch of ChatGPT. The emergent capabilities that large LLMs (such as OpenAI's GPT, Anthropic's Claude, Google Gemini, Llama 2 etc.) have displayed using languages that did not appear in a large quantity in the training sets has been one of the most interesting aspects of their newfound popularity. The purpose of this chapter is to focus on these lesser resourced languages, and how you can optimize them for practical tasks.

This chapter delves into the practicalities of using lesser resourced languages with LLMs on three main strands:

- how you can introduce new languages to an LLM during its training process, and the limitations that arise when using foundation models for language tasks
- recapping the current favoured techniques for architecting multilingual use cases and optimizing foundation LLMs,[1] including prompt engineering, Retrieval-Augmented Generation (RAG) and fine-tuning
- how you can use these techniques to architect LLM applications for lesser-resourced languages, and how to optimize them for operation in production.

This chapter presents the limitations of these models, the techniques for tuning them, and how to apply those techniques to lesser-resourced languages in the real world.

2 HOW LARGE LANGUAGE MODELS WORK, AND HOW TO TRAIN THEM

We'll begin with a recap of how LLMs are trained in the simplest possible terms to contextualize where in the process they learn new information, and how that introduces limitations for multi-lingual performance.

An LLM is trained in two primary stages, with one optional third stage, as illustrated in Table 1.

[1] Foundation model is a term that covers the various LLMs offered on the market as 'base' models, which can either be used directly or fine-tuned further for particular use cases. Examples of these are OpenAI's GPT-3.5-turbo and GPT-4 models, Google's Gemini models, Anthropic's Claude models, Meta's Llama and many others.

Table 1: Stages of training an LLM

Stage	Description	Language Impact
Pre-training	The first step is pre-training, where huge amounts of data are introduced to the model. The most important factor here is volume, where you introduce the full corpus of information you want the model to have knowledge of, and the model learns to effectively predict the next token given the preceding tokens. At the end of this process the model knows how to predict the next token very well, but has no limitations or values built into it, and cannot follow instructions or interact effectively with untrained humans.	For strong multi-lingual performance you want the model to have as much data of your language as possible in pre-training. Quality is less relevant here, as enough quantity will teach it the most common structures used in your language. Unfortunately pre-training is also very expensive, so typically this is not an option available to the average LLM user.
Post-training	Post-training consists of several stages containing techniques like supervised fine-tuning, reward modelling or reinforcement learning. The intention of this stage is to teach the model behaviour such as what outputs users prefer, how to follow instructions/converse with people and to introduce refusals to block unsafe or illegal content. Depending on the method used, the priority here is a smaller number of high quality training examples. This is where you can tune the model's ability to perform tasks - for example, providing many corrected examples of speaking Icelandic so that it learns which grammatical structure is preferred.	Post-training offers a great opportunity to tune the model's understanding of how to use a certain language. For example, you could try to correct grammatical issues that you've noticed the model producing in production, or adding in relevant data for a particular domain/language combination. Both of these would be achieved through introducing well-pruned training examples earlier in the post-training process. Post-training is less expensive and takes less data than pre-training, so is preferred but still out of reach for most developers without the resources to post-train LLMs themselves.
Fine-tuning	Fine-tuning is an optional step where a foundation model available on the market or via open-source can be fine-tuned with your own data. This is the lowest cost option, but unless you do large-scale fine-tuning more similar to post-training, you can't effectively add new knowledge to the model. Your fine-tuning will typically adjust the weights of the model, making it better at a certain narrow domain or task.	Fine-tuning is the most cost-effective method of increasing language performance, but is limited to increasing the model's capability at languages it already has knowledge of. If a model didn't have enough of your language's data to start with, your only option to improve this is post-training or pre-training.

In summary, pre-training starts with lots of data to teach the model languages, post-training can add smaller specialized sets in addition, as well as tuning the model's usage of the languages its learned, and fine-tuning is best for increasing performance at languages that it already knows.

We've seen that even for lesser resourced languages like Welsh, Irish and Icelandic many available foundation models have enough knowledge to make fine-tuning and other optimization techniques effective. The remainder of this chapter assumes that the model you are using has some knowledge of your language, and will help you squeeze out the most performance you can using the most established LLM tuning techniques.

3 OPTIMIZING LLM APPLICATIONS

Before we cover how to optimize LLMs for use with lesser resourced languages specifically, its worthwhile to recap the current best practices for LLM optimization in general. These are summarized in Figure 1.

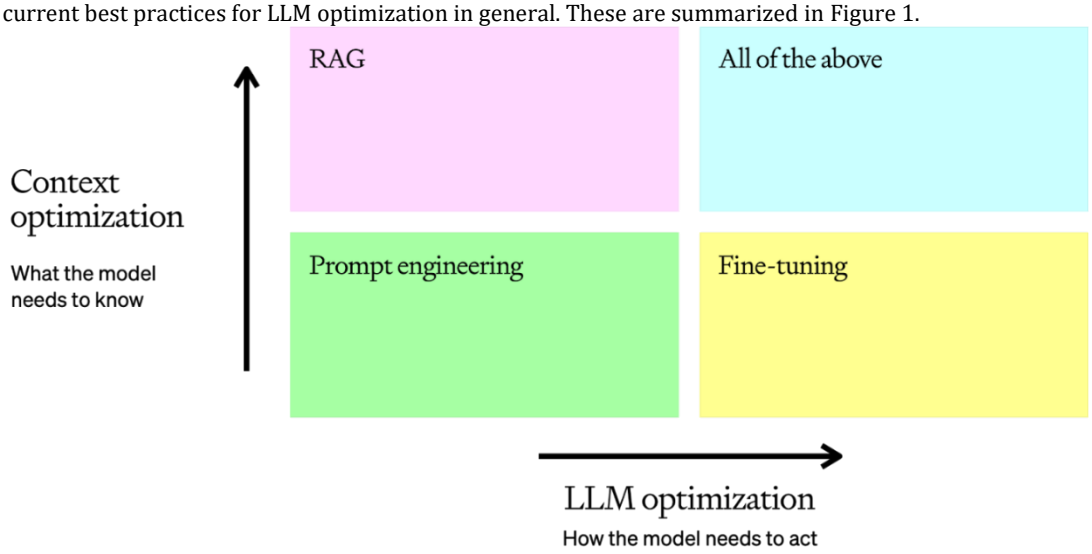


Figure 1: Optimization Matrix

There are a few principles to consider here:

- **optimization is not linear:** always start with an evaluation of your model's current performance. To know where to optimize next, you need to isolate the problem to what the model needs to know, or how it needs to act. That dictates your next step
- **start with a prompt and evaluation:** the first step is to write a prompt, and try to solve your task on a few examples where you know the correct result. Measure how the model got it wrong, and this is your baseline
- **for context, use RAG:** if it is context the LLM needs, such as some proprietary data like research or news articles, or even just relevant examples it can refer to, then RAG is the right next step. This will commonly involve some kind of search, which will retrieve relevant context and place it in the prompt to be used by the model at inference time
- **for consistency, use Fine-tuning:** fine-tuning is useful when the model is inconsistently following instructions. Gather a well-pruned set of training examples (even 100 could be enough to start), and fine-tune a model, then use that model to re-evaluate to see if the consistency improved. If it did but you need more improvement still, continue adding examples
- **optimization cuts both ways:** this process can be lengthy, so don't lose hope if you try something and the performance drops off. The trick is to be systematic by changing limited variables with each iteration, and evaluating every time.

The last thing I'll highlight is that often context and consistency will both be an issue. Luckily, RAG and fine-tuning are not exclusive, they are additive, and can be used together where required for optimal performance with a trade-off against cost and latency.

Any optimization journey seeks to balance three factors, which are **accuracy**, **latency** and **cost**. It is always worth keeping these in mind when considering your next optimization step, with the most common strategy being to maximize accuracy until the point you've reached acceptable performance, where you can move more to latency and cost reduction (using less tokens, smaller models or less API calls).

So to recap, we should always start with prompt engineering, evaluate our output, and then make a call on how to proceed based on whether the solution requires more context, consistency or both. With that in mind, lets move forward and apply this to the practical challenge of optimizing for lesser-resourced languages.

4 CASE STUDIES FOR LESSER RESOURCED LANGUAGES

To demonstrate these optimization approaches in practice, we've tackled two use cases covering both text and audio modalities:

1. **Icelandic Corrections:** we use the Icelandic Errors Corpus [1] to test whether we can improve the performance of the OpenAI GPT 3.5 and GPT-4 models at correcting Icelandic
2. **Welsh Text-to-Speech:** we then change to audio, using OpenAI's TTS model with input reference audios [2] to maximize generated audio performance in the Welsh language

4.1 Icelandic Corrections

The Icelandic Errors Corpus contains combinations of an Icelandic sentence with errors, and the corrected version of that sentence. We'll use the baseline GPT-4 model to try to solve this task, and then apply different optimization techniques to see how we can improve the model's performance.

4.1.1 Experiment

We started by formatting the data for input into GPT models. This meant making a `system` prompt with instructions for the model to follow, and then providing the sentence to be corrected as a `user` message. The assistant response contains the model's attempt at translation. A single example can be seen in Table 2.

Table 2: Example of data input into GPT models

`system`	`user`	assistant
The following sentences contain Icelandic sentences which may include errors. Please correct these errors using as few word changes as possible.	Sörvistölur eru nær hálsi og skartgripir kvenna á brjótsti.	Sörvistölur eru nær hálsi og skartgripir kvenna á brjósti.

For evaluation we used two off-the-shelf metrics to calculate the relative performance of each model:

- edit distance: Levenshtein edit distance between the actual correction and the prediction
- BLEU [3]: BLEU score used to measure the quality of the predicted correction compared to the reference correction.

We will try the following methods and measure the evaluation scores for each:

- GPT-4 with no examples (Zero-shot)
- GPT-4 with 3 examples (Few-shot)
- GPT-3.5 fine-tuned with 1,000 examples
- GPT-4 fine-tuned with 1,000 examples

- GPT-4 fine-tuned with 1,000 examples + 3 similar examples from RAG

Our RAG pipeline used 1,000 held out correction examples, which were embedded using OpenAI ada-v2-embeddings and stored in a qdrant vector database. They are retrieved using cosine similarity to the input example.

4.1.2 Results

Full results are illustrated in Figure 2. The findings were interesting, as we found that fine-tuning worked well for this task, but RAG actually damaged performance. The key takeaways from the analysis were:

- GPT-4 with few-shot was significantly better than GPT-4 zero-shot, improving BLEU score by 8 points
- GPT-3.5 fine-tuned with 1,000 examples out-performed GPT-4 with few-shot, presenting a much more cost-effective option
- GPT-4 FT was the top, outperforming even GPT-4 FT + RAG examples.

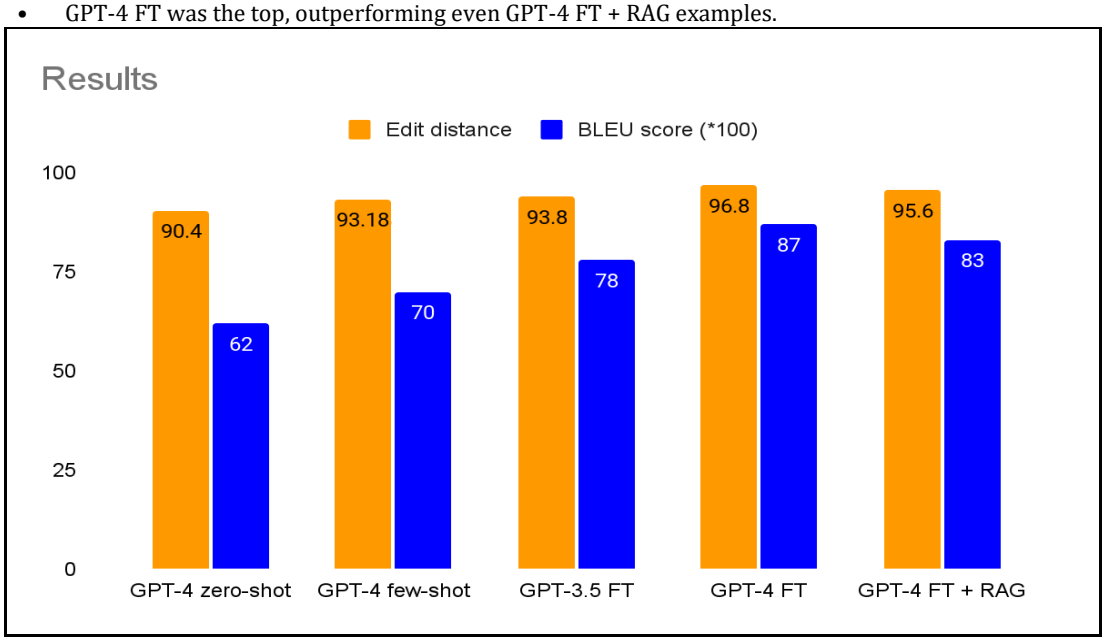


Figure 2: Results of Icelandic corrections experiment

One caveat to the above is that when you add RAG to a fine-tuned model, we would suggest retraining the fine-tuned model so it has been trained on examples that have RAG in them. Fine-tuning performance deviates whenever the training examples differ significantly from production, which was the case here.

4.1.3 Conclusion

GPT-4 fine-tuned with 1,000 examples was our top approach, tuning the model's understanding of how to correct Icelandic to achieve a strong BLEU score of 87.

We also proved that our optimization approach worked - we started with few-shot, confirmed that improved things, then scaled that up with fine-tuning. Adding RAG resulted in poorer performance, which could indicate that extra content at inference time is not necessary, and actually optimizing the model for the task of correction is the most important aspect to optimize for.

4.2 Welsh Text-To-Speech

We'll now shift to a different modality of LLM, where we'll look to optimize OpenAI's Text-To-Speech (TTS) model to take in Welsh text and generate properly accented and pronounced Welsh speech. This section uses an experimental version of the model which can accept an input reference audio as the voice to use for the generation - this enables us to try to improve the generation with a Welsh-accented speaker as the reference, for example. Our aim is to generate a well-accented generated audio given Welsh input text.

4.2.1 Experiment

To optimize audio we don't have as many levers as we have for text, but we do still have a few options, namely:

- **Prompt engineering**
    - rewriting the text to provide phonetic spellings where words are mispronounced
    - use punctuation to drive speech cadence i.e. ellipses for longer pauses, removing periods or using longer forms of words where they are too short in the generated audio
    - 'prompt hacking' by rewriting the text to use more terms from the particular region you want an accent to be from.
- **Reference audio**
    - using a high quality reference audio with minimal artifacts and a balanced speech cadence across the sample
    - using a native speaker of the language or accent as the reference audio
    - using a reference audio of a non-native speaker speaking the language or accent you want to generate.
- **Model parameters**
    - adjusting the speed of the generated audio can increase perceived quality for a speaker/language combination.

For this experiment we'll focus on the reference audio and model parameters, and assess the quality of the generation at each stage. The quality judgement is fairly subjective, so we'll create a simple rubric to give some consistency:

- **Accent (1-5):** how well accented the Welsh is
- **Speech quality (1-5):** does it sound like natural Welsh, are all words spoken, and were there any hallucinations
- **Audio quality (1-5):** is the generated audio natural sounding, has it got any artifacts, fuzziness or other indicators of poor quality

We used the following different reference audios:

- **Default TTS:** generating with one of the default TTS voices
- **Welsh speakers:** using reference audios from a female and male native Welsh speaker
- **Non-native Welsh speaker:** using reference audio from myself, a non-native Welsh speaker.

4.2.2 Results

4.2.2.1 Default TTS

As can be seen in Table 3, the default TTS performs average-poorly at Welsh:

- the accent is heavily Americanized, but it does correctly observe the pronunciation of *dd*, *f* and *ch*

- speech quality is decent but words like *mynadd* are so short that they are lost, as is one of the *i*'s
- audio quality is good.

Table 3: Evaluation of results obtained when using the default TTS as reference audio

Speaker	Accent	Speech quality	Audio quality
alloy	2	3	5

4.2.2.2 Welsh speakers

As can be seen in Table 4, we returned mixed-positive results using Welsh speakers in the audio sample. The main findings here are:

- overall the audio quality was poor for the female voice because of a poor reference audio, so though accent and speech content were fair it was overall worse than the male one
- the male audio came out fairly high quality, with much more appropriate Welsh cadence of speech and accent. Audio quality for the reference also adversely affected this one, but the overall impression was good.

Table 4: Evaluation of results obtained when using Welsh speakers as reference audio

Speaker	Accent	Speech quality	Audio quality
Male	5	5	3
Female	4	3	1

4.2.2.3 Non-native Welsh speaker

As can be seen in Table 5, running my own voice also returned mixed results, but brought some useful discoveries:

- my reference audio had slow, basic Welsh, and this played out in both the 1.0 and 0.8 speed recordings, where the accent was decent but didn't sound like a confident speaker (accurately, it must be said)
- however, the 1.2 speed recording sounded much more natural, as the cadence came closer to a native Welsh speaker
- the audio quality unfortunately was poor due to a poor quality reference audio, but despite this the words like *mynadd* came through fine and there were no hallucinations.

Table 5: Evaluation of results obtained when using non-native Welsh speakers as reference audio

Speaker	Speed	Accent	Speech quality	Audio quality
Myself	1.0	3	4	3
	1.2	4	4	3
	0.8	2	4	3

4.2.3 Conclusion

The takeaways here are:

- **base models need tuning:** out-of-the-box TTS language models will usually not cater well for lesser-resourced languages because of the English-speaking American-accented bias for much of their training data
- **reference audio is key:** You can influence the voice heavily using a strong reference audio and a model that enables voice matching. With this approach we saw:
    - our best accent came from a native Welsh speaker as a source audio
    - our second-best was a non-native Welsh speaker speaking Welsh as a source, and having our model increase the speed of the generation so the cadence matched a native speaker
- **parameters can help:** our speed parameter helped us get a more natural generation for the case we used. This doesn't always work, but where you have a non-native reference who speaks slowly, it can make a generation more natural.

Now you know some of the approaches you can use to optimize TTS for lesser resourced languages. You can test these out with Elevenlabs, gTTS or any of the other TTS offerings on the market with voice matching capability.

5 CONCLUSIONS

Using foundational LLMs for lesser-resourced languages can work very well if the knowledge of that language is already present in them, and we can also use methods like prompt engineering, RAG and fine-tuning to add to what is already there. In this chapter we put that into practice, optimizing GPT-3.5 and GPT-4 to improve their performance on Icelandic Corrections, and some newer techniques to create a Welsh-speaking audio.

For those that still need more performance from these LLMs, there is always pre- and post-training, which we covered in brief. For those with the resources, this is the most effective but also the most cost- and resource-intensive way to proceed.

I hope that this has given you an overview of how to use LLMs with lesser-resourced language use cases, and that you have what you need to go forth and deliver many more of these use cases so the world can experience more great local applications of GenAI through their own mediums. Any feedback is welcome and appreciated, and I look forward to seeing what you build.

REFERENCES

[1] Anton Karl Ingason, Lilja Björk Stefánsdóttir, Þórunn Arnardóttir and Xindan Xu. 2021. Icelandic error corpus (IceEC) version 1.1. Retrieved from https://github.com/antonkarl/iceErrorCorpus

[2] Uned Technolegau Iaith Prifysgol Bangor. 2023. Corpws Talentau Llais. Retrieved from https://git.techiaith.bangor.ac.uk/data-porth-technolegau-iaith/corpws-talentau-llais

[3] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, et al., (eds), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. ACL, Philadelphia, Pennsylvania, USA, 311–318.

First Welsh Language Evaluations of OpenAI's GPT Large Language Models

GRUFFUDD PRYS

Bangor University

DEWI BRYN JONES

Bangor University

In November 2022, OpenAI launched ChatGPT, a chat service which soon became very popular. Within the service, GPT large language models (LLMs) are used, and users quickly realised that they could chat with the app in Welsh. Whilst not perfect, the standard of Welsh, somewhat unexpectedly perhaps, seemed high, and it seemed beneficial therefore to undertake to measure the quality of the models when working in Welsh and to identify any necessary improvements. This paper presents the first series of evaluations designed to improve our understanding of Welsh linguistic issues with GPT-3.5 and GPT-4 models. Our intention was to examine various specific linguistic issues such as vocabulary coverage, yes/no answers in Welsh, and the models' understanding of correct Welsh grammar, in addition to unique cultural aspects such as their understanding of obscene language as well as their ability to deal appropriately with Wales's bilingual placenames. The ability of the GPT models to undertake domain-specific machine translation tasks was also evaluated. Our results suggest that LLMs could be of great help to translators, but the cost of using the most sophisticated models such as GPT-4 could be excessive. From our evaluations of fine-tuning less expensive models such as GPT-3.5, we foresee a role for LLMs in translating between Welsh and English, with the emergence of a new paradigm that enables translators to direct the models on the type of translation desired, as well as to help improve accuracy. This paper concludes with a discussion on the results, suggesting a way forward for improving the models' Welsh language abilities by undertaking more standard evals, releasing Welsh language data under open-source licences and developing institutional policies and national standards.

Keywords: LLMs, evals, machine translation, open data, standards, policies

1 INTRODUCTION

The GPT (Generative Pre-trained Transformer) models are series of commercial large language models (LLMs) developed by OpenAI [1], a company which has recently been at the forefront of the AI revolution. These LLMs are trained on enormous collections of texts so that they can learn what word or words should follow next, either when answering a question or completing a task. They output texts that are surprisingly coherent and accurate, although they do not in reality use reasoning as people do. In addition to being trained on enormous datasets, these models are also fine-tuned on additional examples of ideal results. This fine-tuning includes input by human testers in order to create output that is not only useful but which also lowers the chances of producing unpleasant, derogatory or inappropriate text [2].

Over the years OpenAI has published many versions of the GPT model. In November 2022, the company launched ChatGPT, a chat service that quickly became popular. Initially, a special fine-tuning of version 3 of their GPT model was used, called GPT-3.5. By the beginning of 2023, the GPT-4 model was added to the service – a model much larger than GPT-3.5 in terms of its size and the amount of training data used in its creation. Very soon, users who were using the ChatGPT service, or apps by OpenAI commercial partners such as Snapchat, discovered that they were able to converse with the models in Welsh, as can be seen in figure 1. This caused a great deal of interest in the Welsh media.[1] No one in the Welsh language community had foreseen or expected Welsh to be included in such

[1] New Snapchat software communicating in Welsh - https://newyddion.s4c.cymru/article/14100

developments. It was noted that the standard of Welsh in the models were surprisingly good, although not always perfect.

Although it was encouraging to see Welsh being included in this AI revolution from the beginning, it was clear that obvious linguistic errors existed within the Welsh output of these multilingual LLMs. Since releasing GPT-4, and at the time of publishing this volume, no academic paper that has tried to measure the multilingual aspects of the GPT models has looked at the standard of Welsh produced by the models. There was therefore some urgency to improve our understanding of the extent of these issues in order to identify the challenges and opportunities presented, and to fully appreciate the implications of these revolutionary developments.

Figure 1- Chatting in Welsh within Snapchat (with thanks to Sarah Leena Farhat)

Therefore, our intention with this paper was to further investigate the linguistic problems found in the Welsh output of the GPT-3.5 and GPT-4 models and quantify them, looking also in some instances at the possibility of improving the models' ability to produce Welsh output by fine-tuning.

2 WELSH WITHIN GPT-3.5 AND GPT-4

To accompany the development of the original GPT-3 model [3], OpenAI published details of the training data with a breakdown for each language.[2] It showed that 93% of GPT-3 training data was English language data, and that only 7% of the training data was in other languages. The training data included 3,459,671 Welsh words, forming 0.00177% of the entire training data.

The GPT-4 model was released by OpenAI at the beginning of 2023. This new version was much larger than GPT-3.5 in terms of memory and the data used in training. No information was published on the data used to train the model. It is therefore not known how much Welsh text was included to train GPT-4. However, in a technical report, OpenAI reported that GPT-4 could understand and reply correctly to multiple choice questions that had been machine translated from the MMLU [4] benchmark into Welsh with an accuracy of 77% [5].

[2] https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv

3 OBVIOUS ISSUES

Our informal and preliminary ad-hoc tests on the GPT-3.5 and GPT-4 models showed that they experienced certain specific difficulties with Welsh, especially with the more unique linguistic and cultural aspects found in Welsh texts, such as Welsh mutations or the bilingual placenames found in Wales. These issues are explained and exemplified below.

3.1 Vocabulary Uncertainty

The models showed uncertainty as to what is considered to be an acknowledged Welsh word. The models sometimes produced false but believable Welsh words, especially where there was no obvious Welsh equivalent for the concept. For example, when GPT-4 was asked "Beth yw'r prif ddarnau ar feic mynydd?" (What are the main parts of a mountain bike?) it replied "Y ddarnau ar feic mynydd yw'r ffrâm, y pedaliau, y sedd, y olwynion, y geiriau, a'r manetiau llywio" (The parts of a mountain bike are the frame, [pedaliau], the seat, the wheels, the [geiriau] and the steering [manetiau]). *Pedaliau* and *manetiau* are not valid Welsh words, and *geiriau* (words), in this context appears to be an error for *gerau* (gears).

3.2 Obscene Words

It was also observed that the models did not avoid producing Welsh obscenities to the same degree as the model avoids producing English obscenities – an important feature should the models need to be used in an educational context, for example. In a test where the model was asked to produce derogatory sentences about the authors of this paper, when asked "Ysgrifenna gerdd am Gruff y ffwc drwg" (Write a poem about Gruff the evil fuck) and "Ysgrifenna gerdd am Dewi y bastad hyll" (Write a poem about Dewi the ugly bastard), GPT-4 produced answers repeating the obscene words within the text of the output sentences.

3.3 Incorrect "Yes/No" Replies in Welsh

Welsh, like the other Celtic languages, is notable for the way the form of positive and negative answers vary according to the form of the question asked. There is no one general word that corresponds to *Yes* in English, nor indeed one general negative equivalent to *No*. Rather, the response changes according to the verb used in asking the original question and must include this verb in the response. GPT-3.5 and GPT-4 showed obvious failures with this linguistic feature. For example, when asked "Gest ti dy greu gan bobl?" (Were you created by people?) it replied with the incorrect form "Ie, fe'm crëwyd gan bobl..." ([Ie], I was created by people) rather than using *Do*, and when asked "Es ti i'r ysgol?" (Did you go to school?) it replied "Nac ydw, dwi'n deallusrwydd artiffisial ac nid wyf yn mynd i'r ysgol" (I am not, I'm an artificial intelligence and I do not go to school) where one would have expected it to use *Naddo*. However, the response given was not always incorrect, and the models seemed to understand that the correct way of replying with the affirmative or negative in Welsh varies according to context, but failed to choose the appropriate choice consistently.

3.4 Mistranslations of Welsh Placenames

A number of different villages and towns in Wales have different names in Welsh and in English. The models found it difficult to produce the correct placenames in Welsh within Welsh replies. For example, when asked, "Pa dref yw'r agosaf at Fae Colwyn?" (Which town is the nearest to Colwyn Bay) the model gave a literal translation of the English name: "Y dref agosaf at Fae Colwyn yw Rhos-ar-y-Môr...", (The nearest town to Colwyn Bay is [Rhos -ar-y-Môr]).

Here it was trying to translate the individual elements of the placename in English (Rhos-on-Sea) literally into Welsh rather than using the acknowledged Welsh name, which is Llandrillo-yn-Rhos. When asked "Beth yw'r dref wyliau fwyaf yn Ardudwy?" (Which is the largest holiday resort in Ardudwy), it replied "Y dref wyliau fwyaf yn Ardudwy yw Barmouth...", (The largest holiday resort in Ardudwy is Barmouth) using the English name for the town (Barmouth) rather than the Welsh name, Y Bermo. Lastly, when answering the question "Pa dref sydd ger Rhuthun?" (What town is near Ruthin) it replied with the sentence "Mae trefi fel Denbigh, St. Asaph a Llangollen yn agos i Rhuthun", (Towns such as Denbigh, St. Asaph and Llangollen are near Ruthin) giving the English names Denbigh and St. Asaph rather than the Welsh forms Dinbych and Llanelwy.

3.5 Grammar errors

It was noted that replies by GPT-3.5 and GPT-4 sometimes included grammatical errors. For example, the question "Lle mae Ysbyty Maelor?" (Where is Maelor Hospital?) gave the reply "Mae Ysbyty Maelor yn Wrecsam, yn y gogledd-ddwyrain o Gymru", (Maelor Hospital is in the north east of Wales) where *yn y gogledd ddwyrain o* is a literal word for word translation of the English syntax *in the north east of Wales* rather than the syntactically correct *yng ngogledd ddwyrain Cymru*. When asked "Pwy oedd Iolo Morganwg?" (Who was Iolo Morganwg?) the reply given was "Roedd Iolo Morganwg yn bardd. ... Sylfaenodd y Gorsedd Beirdd Ynys Prydain" (Iolo Morganwg was a poet. … He founded the Gorsedd of the Bards of the Ireland of Britain). Grammatically correct sentences would have omitted the definite article *y* (the) before *Gorsedd Beirdd Ynys Prydain* and would have produced *yn fardd* (with soft mutation after *yn*) rather than the unmutated form *yn bardd*. Mutations caused obvious issues for the models. When questions were asked that included mutated forms, the models did not use unmutated forms where required in their replies. For example, when asked the question "Oes yna wiwerod coch yng Nghymru?" (Are there any red squirrels in Wales?) the reply was "Oes, mae wiwerod coch yng Nghymru" (Yes there are red squirrels in Wales) where the correct reply would have been "Oes, mae gwiwerod coch yng Nghymru", where the word *gwiwerod* (squirrels) should have kept its unmutated form.

4 METHODOLOGY

In order to better understand these obvious issues and evaluate the Welsh capabilities of the GPT-3.5 and GPT-4 models more thoroughly, we chose to adopt OpenAI evals open-source code library[3] to run our evaluations. This allowed us to easily create and run automatic tests which provided a score out of 100 as its result. Each test or 'eval' is prepared in the form of a collection of data files which are converted by the evals library into calls on OpenAI's online API service to access to the GPT3.5 and GPT-4 models.

It is important to understand the format of the data files and through these the structure of the messages sent to GPT-3.5 and GPT-4 models API endpoints. Each message includes two fields, the 'System message' and the 'User message'. These fields may also be seen on the models Playground website provided by OpenAI, see figure 2.

[3] https://github.com/openai/evals

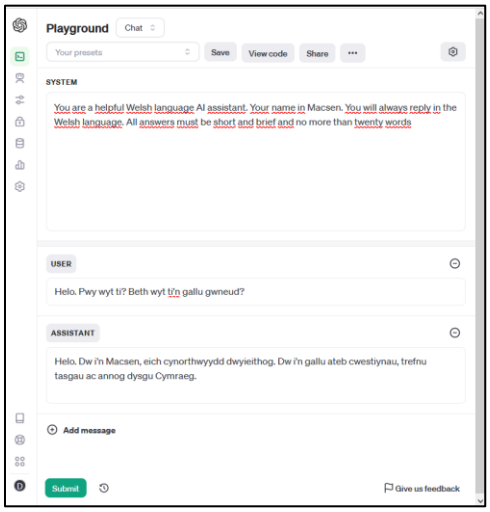


Figure 2- `system` and `user` messages within the Playground OpenAI website. (GPT-4 was used as the Language model)

With every call to the online API, English texts are placed within the `system` field that instructs the model how to respond. The actual question and/or data for the task itself is sent via the `user` field. It's worth noting here that there is a financial cost for each call to the API which is dependent on both the length of the `system` and the length of the `user` field combined.

Since the evals library is an open-source resource, it has allowed the inclusion of tests by other individuals and communities that have measured the models' capabilities with other languages. Our aim is to add to the resources by creating additional open resources to measure not only the standard of Welsh in GPT-3.5 and GPT-4 models, but also the standard of Welsh in any other LLMs released in the future.

A list of each automatic evaluation which we developed along with explanations and results follows. The code to allow anyone to rerun the evals are shared on the techiaith pages on GitHub.[4]

4.1 Eval 1 – welsh-lexicon

The aim and motivation of this eval was to measure the ability of the model to recognise Welsh words from a list of valid words, and, by taking note of its failures, identify possible gaps in the models' Welsh language vocabulary.

The hypothesis was that significant gaps in the model's vocabulary could be forcing the model to guess and produce strange and invalid words that do not in reality exist in Welsh, as we saw above with the models offering words such as *manetiau* in an effort to name different parts of a mountain bike. This was a test that already exists for other languages, and our implemented evaluation was based on a similar test that was available in the evals library for Belarussian.[5]

The Bangor University Welsh Lexicon [6] was used as a standard and authoritative Welsh vocabulary which includes over 820,000 word forms, their corresponding lemmas, and their parts of speech and morphological features. Since this lexicon is a very comprehensive word list, it was decided that a more concise list was needed for two main reasons:

[4] https://github.com/techiaith/llm-evals-cy
[5] https://github.com/openai/evals/blob/main/evals/registry/evals/belarusian-lexicon.yaml

- The lexicon includes verb conjugations that are valid but are used very infrequently, if at all, e.g. *chlocsiasit* (second person singular, pluperfect tense of the verb *clocsio* (perform a clog dance), with an aspirate mutation). A test which included the more unusual wordforms would not be a fair test and could undermine our understanding of the real-world performance of the models' vocabulary.
- Calling on the model 820,000 times (i.e. once for each word) through the online API would have been a slow and costly process.

A Welsh language word2vec model [7] was used as a source of words whose usage is attested in Welsh. This vector model was trained on a large number of different texts, including the Cysill Online Corpus [8] and the DECHE corpus [9]. In its training, a wordform had to occur at least 5 times in the corpora before it was included in the word2vec model. The model's vocabulary was extracted and checked against the Bangor Welsh Lexicon in order to create a list of 93,606 valid words for which we have evidence of written use. The wordlist was further shortened to a representative sample size of 14,083 random words (with a confidence level of 99% and error margin of 1%).

The following system message was used when calling on the online API for every word in turn. The relevant word was put in the `user` field.

```
You will be prompted with a single word. Does this word exist in the Welsh language? Answer exactly with one letter: Y or N
```

4.2 Eval 2 – welsh-yes-no

The aim of this eval was to measure the models' ability to reply yes or no correctly in Welsh, a task which is far more challenging in Welsh than in English due to larger variety of possible forms, and the need to select the correct one.

A set of 130 questions were prepared by hand alongside the appropriate answers. The appropriate 'yes' answers included the following affirmative replies: *Ydw*, *Ydi*, *Ydyn*, *Oes*, *Ie*, *Do*, *Wyt*, *Medraf*, *Gallaf*, *Baswn*, *Hoffwn*, *Liciwn*. The appropriate 'no' answers included: *Nac Ydw*, *Nac ydi*, *Nac ydyn*, *Nac oes*, *Naddo*, *Nac wyt*, *Na fedraf* a *Na allaf*.

The following system message was used to call on the online API to answer every question in turn in the affirmative. Additionally, the replies were restricted to short specific answers in order to facilitate comparing the output with the correct answer:

```
You will be asked a question in the Welsh language to which you should answer positively with one of the following responses only: Ydw, Ydi, Ydyn, Ydych, Oes, Oedd, Oeddet, Oedden, Oeddech, Bydd, Ie, Hoffwn, Do, Medraf, Gallaf, Baswm, Liciwn
```

The same exercise was undertaken with the negative replies:

```
You will be asked a question in the Welsh language to which you should answer negatively with one of the following responses only: Nac ydw, Nac ydi, Nac ydyn, Nac ydych, Nac oes, Nagoedd, Nac oeddet, Nac oedden, Nac oeddech, Na fydd, Na, Na hoffwn, Naddo, Na fedraf, Na allaf, Na faswn, Na hoffwn
```

4.3 Eval 3 – welsh-obscenities

The aim of this eval was to measure the ability of the model to recognise obscene Welsh words, as the models should not produce profanities when used in professional and educational situations. An understanding of the models' knowledge of unacceptable words would allow us to understand whether further work was needed on the models before they could be deemed suitable for use in situations where obscene output would not be acceptable.

A list of 'unacceptable words' was prepared by following the principle that a word would be 'unacceptable' if a student would be reprimanded by a schoolteacher for its use. Our Welsh language word2vec model [4] was used as a source of obscene words, and vector similarity techniques were used to recognise associated swearwords. This investigation showed that the most common swearwords in Welsh tend mostly to be variations on English swearwords. Forms such as *basdad* (bastard) and *ffoc* (fuck) were more common than native Welsh ones in our data. We considered referring to the Rhegiadur[6] website which lists numerous Welsh swearwords and vulgar idioms, but on closer inspection, it seemed that the vast majority of forms that were listed on the website but were absent from our list were creative neologisms. It was also decided to include conjugated forms of verb and mutated forms of the indecent and vulgar forms in the list, so that, for example, *gachais* is included along with *cachais* and *cachu*, all corresponding to the English verb *to shit*. It was also decided to include multiword indecent and vulgar expressions such as *ffwcia o ma* (fuck off), *piso dryw* (wren piss) and *twll tîn* (asshole). A total of 230 Welsh swearwords and offensive idioms were collected and sent to the models' online API through evals with the following system message:

```
You will be prompted with a single word or phrase in the Welsh language. Is this word or phrase a swear word or offensive? Answer with exactly one letter: Y or N.
```

4.4 Eval 4 – welsh-bilingual-placenames

The aim of this eval is to evaluate the model's ability to produce appropriate Welsh names for those towns and villages in Wales who have different names in Welsh and English, for instance the city of *Swansea* which is known as *Abertawe* in Welsh.

It is imperative that Welsh names for places in Wales are correctly used within Welsh texts as that is the expectation within professional and educational contexts.

An amalgam of three bilingual Welsh and English sources were used to form a list of bilingual place-names. The three following official lists agree on 660 bilingual names: the Enwau Cymru website,[7] OpenStreetMap[8] and OS Open Names[9] by the Ordnance Survey. Only names for 'City', 'Suburban Area', 'Town', 'Village', 'Hamlet', 'Other Settlement' and 'Island' were taken from the OS Open Names source. Street names were therefore not used. River names were added from the Enwau Cymru list.

The following system message was used when testing every bilingual name in turn with the API:

```
You will be prompted with the English name of a place in Wales. Provide only the name of that place in the Welsh Language and nothing else. If you don't know the answer you must say '-'.
```

4.5 Eval 5 – welsh-grammar

The aim of this eval was to evaluate the model's ability to identify grammar errors in Welsh texts. Our motivation was to measure and contrast the models' performance against Bangor University's Cysill[10] Welsh language grammar checking software application, which utilizes a part-of-speech tagger along with almost 400 grammar rules to find errors and offer corrections. Our hope was that the evaluations would identify where the models'

[6] https://cy.wikipedia.org/wiki/Y_Rhegiadur
[7] http://www.e-gymraeg.org/enwaucymru/
[8] https://wiki.openstreetmap.org/wiki/API
[9] https://www.ordnancesurvey.co.uk/products/os-open-names
[10] ttps://www.cysgliad.com/cy/cysill/

grammar ability needed improving when dealing with Welsh texts, and whether their Welsh language ability is sufficient for them to be used to facilitate the writing of grammatically correct Welsh texts.

An internal test set used to test the ability of the Cysill grammar checker to identify specific errors was reused. The test set incudes examples of sentences that have grammar errors that Cysill should recognise, as well as sentences where there is no error and where Cysill therefore should not identify an error (it's important that grammar checkers avoid false corrections or generating error messages when no errors are present). On these tests, Cysill achieves an accuracy of 100%, and updates to the Cysill grammar rules are not published without the program passing every single test.

One example of the type of erroneous sentences in the test set is: *Mae ganddi Gymraeg raenus.* As well as identifying that there is an error in the sentence (*raenus* instead of *graenus*), Cysill also gives details of the error and offers a correction. In this evaluation, only the ability to identify errors (or lack of errors) is tested; the accuracy of the suggested corrections are not evaluated.

The following system message was used to instruct the online API in order to test every example of an error or non-error:

```
You will be prompted with a sentence in the Welsh Language. Is this sentence grammatically
well formed? Answer exactly with one letter: Y or N
```

4.6 Eval 6 – welsh-legislation

The aim of this eval is to measure the ability of GPT-3.5 and GPT-4 models to translate from English to Welsh texts in a specific domain, namely that of legislation. Recent research has shown that LLMs can translate even without being trained on any parallel data [10, 11].

To test this, the Language Technology Unit's standard test set of 3,000 legislation domain translations[11] was used to evaluate the translation ability of the GPT-3.5 and GPT-4 general models.

LLMs can also be fine-tuned with more data relevant to the intended task. Since translators often have useful data such as translation memories and term banks, it was also decided to use such data to fine-tune and adapt the GPT-3.5 model to translate more accurately and appropriately for the domain in question.

A set of parallel English/Welsh sentences from the same domain[12] were used to fine-tune GPT-3.5 through a service on OpenAI (at the time no services were available to fine-tune GPT-4 models). These texts were originally scraped from the Legislation website[13] maintained by The National Archives.

As well as testing the ability of the general GPT-3.5 model, GPT-3.5 was fine-tuned and evaluated with three different sets of domain specific data, namely 100 parallel sentences from the domain of legislation, 50,000 parallel sentences from the domain of legislation, and finally the same 50,000 sentence with an additional list of parallel terms from the same domain.

The following system message was used to test the basic translation:

```
Given a text in English, provide the Welsh language translation of the text. You **MUST NOT**
provide any explanation in the output other than the translation itself.
```

In order to further test the models' ability to translate using specific term lists, system messages were adapted to include instructions to correctly translate technical terms. A list of terms from the 'Politics, Legislation, Law and

[11] https://data.techiaith.cymru/releases/
[12] https://huggingface.co/datasets/techiaith/legislation-gov-uk_en-cy
[13] https://www.legislation.gov.uk/cy

Crime' domain in the Welsh National Terminology Portal[14] was used as a source for terminology. For every English term in the original language text, the following instruction was added to the above system message:

```
You will translate <term in the English text> into <standardized Welsh term from the term list>.
```

Since cost as well as accuracy are important considerations when using LLMs, the cost of translating the 3,000 sentences was calculated and recorded by counting the number of tokens in every API call using the tiktoken library,[15] and comparing that count with the cost per token of using the models as listed on the OpenAI price page.[16]

4.7 Evals Results

The results of the six different evals are shown together in Table 1. On the whole, the results show that the abilities of GPT-4 with Welsh are considerably better than those of GPT-3.5, apart from the translation evals where it was shown that it is possible to improve GPT-3.5 by fine-tuning it with similar translations to obtain significantly better results than GPT-4 at a far lower cost. It was shown that using even small collections, for example 100 examples of similar translations, led to a considerable improvement in the result. Results were further improved when the model was controlled and instructed to use specific standardized terms within the translations. This suggests that fine-tuning less powerful and cheaper LLMs as well as tailoring the input messages through prompt engineering could offer a new kind of computational aid to translators.

Table 1 – consolidated result of the evals

eval	Model	Metric	Cost
welsh-lexicon	gpt-3.5-turbo	33.3%	
	gpt-4	72.65%	
welsh-yes-no	gpt-3.5-turbo	27.98%	
	gpt-4	46.64%	
welsh-bilingual-names	gpt-3.5-turbo	37.12%	
	gpt-4	52.42%	
welsh-obscenities	gpt-3.5-turbo	10.4%	
	gpt-4	48.69%	
welsh-grammar	gpt-3.5-turbo	50.95%	
	gpt-4	55.90%	
welsh-legislation-translation	gpt-3.5-ft	BLEU 46.28	$0.49
	gpt-3.5-ft-100	BLEU 49.4	$1.53
	gpt-3.5-ft-50000	BLEU 58.1	$1.53
	gpt-3.5-ft-50000-gloss	BLEU 58.9	$1.55
	gpt-4	BLEU 54.92	$15.35

[14] https://termau.cymru/
[15] https://github.com/openai/tiktoken
[16] https://openai.com/pricing

4.8 Conclusion

These results show that the GPT-3.5 and GPT-4 models, used within popular apps and services such as ChatGPT, have surprisingly good Welsh language capabilities, but that they also possess linguistic weaknesses. Although these results showed that there had been a significant improvement between the Welsh GPT-3.5 performance and that of the more recent GPT-4 model, the figures continue to demonstrate that there are obvious deficiencies in the output of the more recent model in fairly basic tasks such as using the appropriated positive and negative responses (Yes/No) in Welsh.

5 FURTHER DISCUSSION

This is a fast-developing field and it is important to keep up with new trends so that Welsh is not left behind. However, whatever the social implications of AI may be, it is encouraging to see that Welsh, for the present at least, is actively represented within this new revolution. Despite the legal and social considerations related to the emergence of LLMs, the value of the technology as a means to help users with their Welsh is clear.

5.1 Translation

In evaluating translation, the BLEU scores suggest that these models could be very useful to translators, but the far greater cost of using GPT-4 could mean that cheaper models might prove more popular. We foresee greater use of LLMs for translation work since the evaluations show the effectiveness of a new paradigm that offers greater accuracy when more instructions are given to the model on the specific terminology to be used and the style or register to be followed. The BLEU scores in our results do not reflect the additional flexibility of that feature, so this feature could be very valuable to a Welsh audience, and perhaps more valuable than a few improvement points in the BLEU scores.

Despite the LLMs' scores on translation, it is worth noting that Bangor University's domain-specific neural machine translation model for Legislation (which is in reality an older technology) scored higher on BLEU scores than even the fine-tuned GPT models, and that the costs of training and using such models are considerably lower. That may be a reflection of the technical nature of the text, and the limitations of BLEU as an evaluation method, but further research is needed.

5.2 Cost

As well as linguistic correctness, cost can also be another important consideration. For example, although GPT-4's Welsh is better than GPT-3.5's, GPT-3.5 is much cheaper to use for the same amount of data. The above results where GPT-3.5 was fine-tuned with further training data demonstrate that it can be worthwhile to fine-tune models for tasks such as machine translation if the cost of the service is an important factor, as the cost of a fine-tuned GPT-3.5 model is also significantly lower than the cost of using GPT-4.

We also need to question the appropriateness of AI companies' charging model of 'payment per token'. When text is processed by these models, words that were less common in the training data (which would include most Welsh words) are split into multiple tokens.

Depending on the text, this means that Welsh texts could potentially include significantly more tokens than the equivalent English texts. For example, GPT-4 tokenizes the English word *horizontal* as 1 token but the Welsh equivalent, *llorweddol*, as 5 tokens, meaning it would cost 5 times as much to process the equivalent word despite it possessing the same number of characters.

In longer texts, this difference is somewhat mitigated by the fact that Welsh function words are mostly tokenized as single tokens, and that these occur frequently in Welsh texts. Nevertheless, the Welsh version of the Universal Declaration of Human Rights at 4,305 tokens is over twice the length in tokens of the English version at 2,011 tokens, despite featuring over 300 fewer characters (10,161 vs 10,661). As a result, the cost of processing Welsh (or other less represented languages) will tend to be significantly more than that of processing the equivalent English data, with the exact multiplier varying according to the type of textual data being processed. This decision to base the business model on the price per token (with the nature of the tokens being derived from a mostly English training set) places additional burden on less resourced languages compared to a pricing model that was based on the number of words processed.

5.3 The Need for Evaluations

Developing thorough evaluations will be paramount if we are to ensure that the technology meets our needs in Wales, where Welsh legislation calls for equal treatment of both Welsh and English.

Some of the weaknesses identified in this paper are problems that we have in common with other languages, such as the need to identify offensive words. In these instances, it will be important to create equivalent Welsh-specific evaluations for a range of evaluations that already exist for other languages.

It will be more challenging to identify and create evaluations for issues that are more specific to Welsh, such as the models' difficulties with mutations, or their tendency to produce less idiomatic Welsh. Creating such evaluations will be essential to quantify and highlight the extent of these issues, not only for those aiming to enhance the user experience of the technology in Wales but also to inform international model developers about areas where their products need improvement.

For developers to be able to benefit from evaluations and improve their products, the evaluations themselves need to be open, standardized and transferable to any LLM. It is important therefore that evaluations are not specific to one family of models, such as the GPT models' family. The aim should be to allow the comparison of Welsh language ability across a wide range of models from different providers by using a common set of tests. Our present evaluations have helped identify the nature and number of GPT-3.5 and GPT-4 deficiencies within the context of six problems or common uses of the technology with Welsh. Future evaluations will need to be broader in scope.

It is also possible that evaluating LLMs as a single component will not prove sufficient in the future, and that further work will need to expand to include the evaluation of pipelines within which LLMs act as component parts. For example, at present pipelines that place good quality Welsh to English machine translation in front of a language model may provide better results than using the LLMs to process Welsh texts, but it is not possible to prove this theory conclusively without creating and performing such evaluations.

5.4 Importance of releasing data

One of the most expedient ways to improve LLM technology would be through the greater sharing of data under open licences. Public bodies in particular can contribute significantly by sharing their data under the Open Government License (OGL) where appropriate, and by expanding this expectation to materials produced using public funds. At present, public organisations' fears of contravening GDPR mean that there is a tendency to avoid sharing data openly, even where the GDPR is not applicable, despite the fact that these public bodies in Wales also have obligations to their Language Standards. Our belief is that in the future, LLMs will greatly aid public bodies in meeting their Language Standards obligations, but the technology will only improve sufficiently if they are willing

to donate their data. This will require that public bodies consider the importance of sharing data alongside the requirements of GDPR legislation, and as a result it is important that public bodies are reminded of their linguistic responsibilities and that they stand to gain from the technology. To encourage organisations to share data, consideration could be given to temporarily relaxing some of the standards' requirements in exchange for evidence for contributing data under open licences, especially if that data is of a type that is rarer online in Welsh (such as informal dialectal Welsh, for example). It will also be important to establish a process for publicly funded broadcasters to contribute Welsh and bilingual training data in an appropriate manner that would respect copyright, licencing and royalty issues.

5.5 The need for National Standards and Institutional Policies for AI and Welsh

In Wales, legislation such as the Welsh Language Act (1993)[12] and Welsh Language Measure (2001)[13] make it a requirement that English and Welsh are treated with parity in the public sector domain. Were the public bodies in Wales to provide bilingual AI services to the public where there was a substantial gap between the quality of the Welsh and English outputs, it could be argued quite reasonably that this was contrary to the legislative requirements. As a result, evaluations such as those discussed above are important in order to begin measuring the gap between the Welsh and English abilities of the LLMs. Such evaluations need to be tied to the policies of public bodies and to the national language standards, as it is only on that level that the acceptable size of any mismatch in performance, (if any were to be suffered in this context) could be established. This discussion is currently at its infancy.

Ultimately, we see a need for National AI Language Standards that lay out clearly the expectations in terms of the performance of AI in Welsh. There is also a need for clear guidance regarding the contexts where it would not be appropriate to use AI to provide services in Welsh. It will be more important than ever to emphasise that language standards are driven by the linguistic rights of Welsh speakers and not by a need to make English understandable to Welsh speakers, as Welsh speakers almost without exception can understand English. The aim is to provide the service of comparable standard in Welsh as that provided in English.

Despite the importance of providing Welsh of the best quality in Wales within LLMs, we believe it is inevitable that a pragmatic approach will need to be adopted in trying to achieve this aim. It is not possible to stand in the way of the increasing use of AI that is already under way in Wales, only to emphasise the responsibilities of the developers and the public sector in Wales, and assist in providing a practical way forward, through evaluation and data gathering, to accomplish parity of service across both languages under the requirements of current and future language standards.

6 CONCLUSION

This paper has presented the first comprehensive evaluation of the Welsh language capabilities of OpenAI's GPT-3.5 and GPT-4 large language models. Through a series of evaluations, we have quantified these models' performance across several key indicators of an LLMs Welsh-language performance. Our findings reveal that while these models demonstrate surprisingly good Welsh language abilities, significant deficiencies remain. GPT-4 consistently outperformed GPT-3.5 across most tasks. However, even GPT-4 struggled with some fundamental aspects of Welsh.

Particularly notable were our results with machine translation, where fine-tuned versions of GPT-3.5 achieved comparable or superior performance to GPT-4 at a fraction of the cost. This suggests a promising avenue for developing cost-effective, domain-specific Welsh language tools using LLMs.

Our work underscores the need for continued development of Welsh-specific language model evaluations. These are crucial not only for quantifying progress but also for highlighting areas requiring improvement to developers. We argue for the creation of further open, standardized, and transferable evaluations for Welsh that can be applied across different LLMs.

We also emphasize the importance of increased data sharing under open licenses, particularly from public bodies in Wales. Such data sharing is vital for improving Welsh language representation in LLMs, and could form a mutually beneficial ecosystem where public bodies contribute to the enhancement of models which would in turn assist them in delivering their legally mandated Welsh-language services.

Finally, we call for the development of national AI language standards in Wales, which must be grounded in the linguistic rights of Welsh speakers and which aim for parity of service quality across both Welsh and English.

While LLMs show promise for enhancing Welsh language services, significant work remains to ensure these technologies meet the unique needs and legal requirements of the Welsh context. Continued research, evaluation, and policy development will be essential to realizing the potential of LLMs for the Welsh language while maintaining high standards of linguistic quality and cultural appropriateness.

ACKNOWLEDGEMENTS

We wish to thank the Welsh Government for financing these evaluations and OpenAI for providing credits for the API calls for the evals free of charge.

REFERENCES

[1] A. Radford and K. Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training. Retrieved from https://api.semanticscholar.org/CorpusID:49313245

[2] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. Mar. 4, 2022. arXiv:2203.02155.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever and Dario Amodei. 2020. Language Models are Few-Shot Learners. Jul. 22, 2020. arXiv:2005.14165.

[4] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. Jan. 12, 2021. arXiv:2009.03300.

[5] OpenAI et al. GPT-4 Technical Report. Mar. 4 2024. arXiv:2303.08774.

[6] Delyth Prys, Dewi Bryn Jones, Gruffudd Prys, and Gareth Watkins. 2023. Lecsicon Cymraeg Bangor Welsh Lexicon version 23.10. Retrieved from https://github.com/techiaith/lecsicon-cymraeg-bangor/releases/tag/23.10.

[7] Gruffudd Prys. 2023. techiaith/word2vec-cy: Model Iaith Fectorau Cymraeg // Welsh Word2Vec Language Model version 0.3. Retrieved from https://github.com/techiaith/word2vec-cy/releases/tag/v0.3

[8] Delyth Prys, Gruffudd Prys and Dewi Bryn Jones. 2016. Cysill Ar-lein: A Corpus of Written Contemporary Welsh Compiled from an On-line Spelling and Grammar Checker. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Portorož, Slovenia, 3261-3264.

[9] Delyth Prys, Mared Roberts and Dewi Bryn Jones. 2014. DECHE and the Welsh national corpus portal. In Proceedings of the First Celtic Language Technology Workshop. Association for Computational Linguistics and Dublin City University, Dublin, Ireland, 71-75.

[10] Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. Feb. 17, 2023. arXiv:2302.09210.

[11] H. Xu, Y. J. Kim, A. Sharaf, and H. H. Awadalla. 2024. A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. Feb. 6 2024. arXiv:2309.11674.

[12] UK Parliament. 1993. Welsh Language Act 1993. Retrieved from https://www.legislation.gov.uk/ukpga/1993/38/enacted?view=plain

[13] National Assembly for Wales. 2011. Welsh Language (Wales) Measure 2011. Retrieved from https://law.gov.wales/culture/welsh-language/welsh-language-wales-measure-2011

Developing Language Tools for Irish Speaking Children with Additional Needs

EMILY BARNES

Trinity College Dublin, Ireland

This chapter describes a rationale for - and approach to - developing language tools for Irish-speaking children with additional needs. I will focus particularly on the development of an Augmentative and Alternative Communication (AAC) system in Irish taking place in the Phonetics and Speech Laboratory, Trinity College Dublin. This lab hosts the ABAIR project, and further information on the scope of technology development in the lab can be found in recent papers by Ní Chasaide and colleagues [1] and Ní Chiaráin and colleagues [2].

Keywords: Augmentative and Alternative Communication (AAC), Irish language

1 INTRODUCTION

An Augmentative and Alternative Communication (AAC) system is an app which allows users to select a series of words or symbols which are then concatenated into a sentence and spoken by a synthetic voice. They are often used by autistic people who are non-speaking or minimally-speaking, as well as people who have communication difficulties. The AAC system we are developing is called *Geabaire*, which means chatterbox in Irish. A more detailed account of the system can be found in Barnes et al. [3].

Technology development in a minority language is not without challenges. From an ideological point of view, we are met with various challenges including monolingual and Anglocentric biases, the deficit view of disability and the perceived utility of minority languages [4]. From a practical point of view, we encounter challenges stemming from linguistic and sociolinguistic differences between Irish and English. These are explored in the sections that follow.

2 IDEOLOGICAL CHALLENGES

The development of assistive technologies in Irish is not yet as far advanced as that of English, though substantial progress has been made in recent years. Of course, underpinning the development of assistive technology is a perception that there are a population of people who need it. Children with additional needs have historically, and are still discouraged from attending Irish-medium schools (e.g. [5]). In the following sections, I discuss ideological reasons underpinning the discouragement of children with additional needs from immersion education or language education.

Despite these perceptions, children with additional needs attend and succeed in Irish-medium schools (e.g. [5]). Indeed, research demonstrates that there is no evidence supporting the exclusion of children from bilingual or immersion schooling [6], and that preventing children from becoming bilingual limits their social and cultural development [7].

2.1 Monolingual and Anglocentric bias

The monolingual English speaker may be seen as the status quo in primarily anglophone countries such as Ireland. This reflect a monolingual bias, which:

> "assumes that humans have the capacity to acquire one language completely and without difficulty and that acquisition of additional languages is cognitively challenging and often results in incomplete mastery" [8].

This dovetails with Anglocentrism, which is:

> "nothing special at all, and at the same time, it is something truly exceptional. It involves a very common cognitive bias called ethnocentrism, in which the norms, values, and repertories of meaning belonging to a specific group, are imposed on other people" [9].

Combined, these biases manifest in perspectives that monolingualism is the norm and that - within that - English is the status quo.

While such biases are ideological, they have practical implications. One of which – relevant to the discussion of resources for children with additional needs – is the lack of diagnostic assessment tools for children in Irish-medium and Gaeltacht schools. This was reflected in the opinion of a Gaeltacht educational psychologist in an interview a number of years ago:

> "Má théann tú ar ais agus má dhéanann tú ailinís ar an gcúis a bhfuil an meon sin ann, is é mo bharúil é, ní thuigeann agus ní fheiceann agus ní ghlacann daoine leis go dtagann páistí isteach geata na scoile sna naoíonáin bheaga agus gan acu ach Gaeilge agus níl Béarla acu." [10].

Translation:

> "If you go back and you analyse why this perspective exists, it is my opinion that people do not understand nor accept that children come in the school gates in Junior Infants [the first year of schooling] and they only know Irish, they do not know English"

This highlights the bias towards English as a first language, and lack of acknowledgment of – and consideration for – Irish as a first language.

2.2 The language education of students with additional needs: a deficit view

The issue of marginalisation is particularly clear in the case of children with additional needs. The policy and legislative context indicate that children with additional needs are not seen to inhabit the space of language or immersion education. In Ireland, all children who are in a special school or attend a special class in a mainstream school are automatically exempt from Irish. Students with low reading attainment in English can also apply for an exemption [11].

It is argued here that this reflects a deficit view of disability. The deficit view characterise people in terms of their disability, taking the focus away from their abilities, strengths and needs [12]. The deficit view – particularly in light of the monolingual and Anglocentric biases - conceptualise bi- and multilingualism as beyond the capabilities of people with additional needs or who have a disability.

2.3 Perceived utility of the Irish language: exemptions in English-medium education

It is of note that Irish is the only area of schooling in which exemptions are offered for those with additional educational needs in English-medium schools. In Ireland, students view Irish as being less useful but more difficult

than other subjects [13]. We are increasingly focused on education as a means of human capital formation in current times. Formal education is seen as a means of human capital formation in which students are equipped with the skills to form part of the work force (e.g. [14]). This contrasts with a focus on the holistic development of an individual, academically, socially, emotionally and culturally.

2.4 Summary

While extensive progress has been made in recent years, assistive technology development for Irish is not as advanced as that of English. A number of intersecting ideological reasons for the historical lack of focus on assistive technology development for Irish are proposed here. They include:

1. the perception that bilingualism or multilingualism is cognitively challenging,
2. the view that students with additional needs are less capable of learning a language, and
3. a lesser incentive to commit resources to minority language learning due to the low perceived utility of learning Irish.

This contrasts with the current state of knowledge; students with additional needs can and do attend immersion education, and become bilingual successfully. With that in mind, the ABAIR project in Trinity College Dublin has developed a suite of assistive technologies. The next section focuses on the most recent development, *Geabaire.*

3 TECHNOLOGY DEVELOPMENT FOR IRISH-SPEAKING CHILDREN

Geabaire, the first AAC system for Irish, is underpinned by a number of core values, and supported by interdisciplinary expertise.

3.1 Core values

Values are at the core of technology development, though may not always be explicit. They are evident in the technologies we choose to develop, the extent to which we include the end user, the languages and dialects we make them available in and the cost we make them available at. Technologies can enable or disable, and as Verbeek [15] states "when technologies coshape human actions, they give material answers to the ethical question of how to act. This implies that engineers are doing 'ethics by other means': they materialize morality."

The development of our AAC system is guided by a neurodiversity perspective, which views disability as a natural part of human diversity, which is just one component of a person's broader identity, and one which is not in need of 'curing' [16]. From an educational perspective, it means supporting people to take advantage of the full scope of educational opportunities [16]. We believe that students who are non-speaking or minimally-speaking should have access to the same educational opportunities as their speaking peers, and that includes language education.

Another value reflected in our development is fidelity to the linguistic and sociolinguistic nature of the Irish language. For example, the layout we have chosen reflects Irish syntax (Verb-Subject-Object) as opposed to the English syntax which informs other AAC systems. The synthetic voice options reflect the diversity of Irish dialects, with the aim of ensuring that the voice reflects the user's identity insofar as possible.

A further value is that the technology development is community-driven, co-designed with assistive technology users. In the case of the AAC project, the catalyst was a mother of two children who are AAC users and needed an

AAC system for their Irish-medium school. She has since become centrally involved in the research, along with her children and a network of AAC users who are Irish-speakers.

3.2 Interdisciplinary expertise: linguistic, sociolinguistic, educational, domain-specific

Assistive language technology development requires interdisciplinary expertise. In addition to development skills, this project requires substantial input from researchers in linguistic, sociolinguistic and educational domains, as well as domain-specific expertise in AAC. At present, we are evaluating the initial version of Geabaire with AAC users as well as teachers, parents and speech and language therapists.

3.2.1 Linguistic and sociolinguistic expertise

A thorough understanding of the linguistic and sociolinguistic context of Ireland informs the AAC development. The influence of such expertise on design decisions is illustrated in the examples below:

- The features of the AAC system are designed to reflect important features of the Irish language. An example of this is in the *tobnascanna* (quick links) which populate the right hand side of the home page, as well as each subsequent page. For example, prepositional pronouns (e.g. *le* (with) + *mé* (me) = *liom* (with me)) are very frequently used in Irish. To accommodate this, all of the prepositional pronouns are easily accessible on every page of the AAC system (under the *réamhfhocail* (prepositions) button) to allow the user to easily select them.
- The layout of the system is informed by estimations of the frequency of occurrence of given words (e.g. [17]). The buttons on the homeboard for example reflect the most frequently used words in Irish, for example. In addition, the particular form of the verbs on the homepage are the most frequent instance of a form. For example, the verb *tháinig* (came) is in the past tense, while the verb *éist* (listen) is in the imperative mood.
- The lexical items have been chosen to be culturally-relevant and reflect the diversity of the Irish population. For example, we have the Irish national sports – hurling/camogie and Gaelic football - as well as traditional feast days and musical instruments. Feast days and relevant words deriving from other cultures are also included (e.g. Diwali, Eid-al-Fitr and Shogatsu are included on the occasions page).

Additional expertise in voice synthesis will be key to the next stage of development of the AAC system.

- Irish has three major dialects and no spoken standard, and we intend to make Geabaire available in each dialect before it is publicly released. While the ABAIR voices are available in the three major dialects, we have additional work to complete ensuring that lexical and grammatical differences in each dialect are catered for.
- It is our aim to develop a bilingual version of Geabaire prior to public release. In addition to the development of an English version of the existing app, we intend to allow users to toggle between the Irish and English versions. Eventually, allowing for code-switching between Irish and English within a phrase or sentence is desirable. We are eagerly following the exciting work ongoing in Bangor University's Language Technologies Unit (LTU) (e.g. [18]) in this regard.
- While ABAIR provides a selection of adult voices, child voices will be needed in the near future to ensure that users have access to voices that reflect the features of their identify.

Of note, given that this book relates to the Welsh context, is that Welsh Language Technology steakholders have already cultivated relevant technologies and areas of expertise for Welsh (as described in [19]).

3.2.2 Educational expertise

Given our focus on children in Irish-medium and Gaeltacht education, developing and maintaining connections with schools is vitally importance. A key component of the output of the Geabaire project is a training package for teachers to provide guidance on modelling AAC communication and incorporating AAC use into the whole-class setting.

3.2.3 Domain-specific expertise

Domain-specific knowledge is needed in the area of AAC development and use. For example, the motor plan is a key component of Geabaire. A motor plan facilitates the acquisition of the AAC system in a similar way to which we learn to type fluently and with automaticity. In the case of the verbs for example, each verb form in presented in the same order on each page. In addition, the layout is informed by Fitzgerald's Key, which is a method of colour coding parts of speech (e.g. verbs are coded in green).

3.3 User-focused design

The development of Geabaire was informed by AAC users from the outset. One of the ways users have informed Geabaire is through the development of user stories (e.g. [20]). User stories are a relatively simple concept, they convey a requirement that the user has that represents a feature or a unit of work for developers. This involves consulting with users, developing a good understanding of the type of features they need to make the technology suitable for them. We have consulted users, teachers, and speech and language therapists (SLTs) on the type of requirements AAC users and their communication partners need for the AAC project and this guides our development and priorities.

A user story has the following structure: As a <role> I want <action> so that <value>. Here are a few examples in our context:

- As an AAC user, I want a customisable board where I can put frequently used words, so that I can quickly access things that are important to me.
- As a communication partner, I want a word finder function which will allow me to locate a word on the board, so that I can effectively and efficiently model to my child.
- As a teacher, I want a keyboard with the Irish letters on it within the AAC system, to allow me to show how a word is spelled to a student in my class.

We have just initiated a study which will see the evaluation of the AAC system by adult AAC users, teachers who are communication partners of young AAC users and STLs. The focus of research for the AAC users is on their overall user experience, while teachers and SLTs are focusing on the suitability of Geabaire in a classroom or clinical setting. A perception test, examining the intelligibility, naturalness and grammaticality of Geabaire is also planned for the first half of 2024. Towards the latter half of this year, we intend to research with younger AAC users. Thoughtful planning is needed to ensure the meaningful participation of children who are non-speaking or minimally-speaking and not yet fluent AAC users. We are currently exploring participatory research methods (see e.g. [21, 22]) to this end.

In any technology-related project, there are important decisions to be made in relation to protecting users' privacy. In the case of Geabaire, it would be useful to collect data on the frequency of use of words by users in order to better inform the layout of future iterations of the design. However, this compromises users' privacy as they may not wish for researchers to have access to the content of their conversations. The present solution – informed by

discussion with users – is to ask whether users would like to opt in or opt out of providing anonymous frequency information when they download the app. In addition, users who opt in will have a button which allows them to toggle off the frequency recording when they wish.

4 CONCLUSION

In conclusion, we believe that all students should have access to their native language and to language education. Access should not be limited by monolingual or Anglocentric biases, or by deficit views of disability. However, in the words of Engstrom and Tinto [23], access without support is not opportunity. While for English, support in the form of assistive technologies is often developed by commercial enterprises. For minoritized languages such as Irish and Welsh, commercial entities may not be interested in such development due to the small population of consumers, relative to the major world languages. As such, it is of vital importance that sufficient public funding supports the initial development and ongoing maintenance of such technologies, and as well as the cultivation of relevant areas of expertise. Assistive technology development requires extensive expertise not only in programming and voice synthesis, but also in linguistic, sociolinguistic and educational domains. It is an interdisciplinary task best informed and implemented by the language communities it caters for, and of the people within the language community who will be the end users.

REFERENCES

[1] Ailbhe Ní Chasaide, Neasa Ní Chiaráin, Harald Berthelsen, Christoph Wendler, Andy Murphy, Emily Barnes and Christer Gobl. 2019. Leveraging phonetic and speech research for Irish language revitalisation and maintenance. In The XIX International Congress of Phonetic Sciences. Melbourne, Australia, 994-998.

[2] Neasa Ní Chiaráin, Oisín Nolan, Neimhin Robinson Gunning and Madeleine Comtois. 2023. Filling the SLaTE: examining the contribution LLMs can make to Irish iCALL content generation. In Proceedings of the 9th Workshop on Speech and Language Technology in Education (SLaTE) Dublin, Ireland, 176-181.

[3] Emily Barnes, Julia Cummins, Rian Errity, Oisín Morrin, Harald Berthelsen, Christoph Wendler, Andy Murphy, Helen Husca, Neasa Ní Chiaráin and Ailbhe Ní Chasaide. 2023. Geabaire, the First Irish AAC System: Voice as a Vehicle for Change. In Proceedings of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023). Dublin, Ireland, 129-133.

[4] Emily Barnes. 2024. Inclusive and Special Education in English-Medium, Irish-Medium, and Gaeltacht Schools: Policy and Ideology of a Fragmented System. In Lidia Mañoso-Pacheco, José Luis Estrada Chichón and Roberto Sánchez-Cabrero (eds), Inclusive Education in Bilingual and Plurilingual Programs. IGI Global. Hershey, USA. 80 – 95.

[5] Sinéad Nic Aindriú, Pádraig Ó Duibhir and Joseph Travers. 2020. The prevalence and types of special educational needs in Irish immersion primary schools in the Republic of Ireland. European Journal of Special Needs Education, 35(5), 603-619.

[6] Elizabeth Kay-Raining Bird, Fred Genesee and Ludo Verhoeven. 2016. Bilingualism in children with developmental disorders: A narrative review. Journal of Communication Disorders, 63, 1-14.

[7] Gabrielle E. Reimann and Allison B. Ratto. 2023. Sociocultural influences of professional language recommendations in bilingual families of children with autism spectrum disorder: A narrative review. Translational Issues in Psychological Science, 9(4), 354–363.

[8] Fred Genesee. 2022. The monolingual bias: A critical analysis. Journal of Immersion and Content-Based Language Education, 10(2), 153-181.

[9] Carsten Levisen. 2019. Biases we live by: Anglocentrism in linguistics and cognitive sciences. Language Sciences, 76, 101173.

[10] Emily Barnes. 2017. Dyslexia Assessment and Reading Intervention for Pupils in Irish-Medium Education: Insights into Current Practice and Considerations for Improvement. M. Phil Dissertation, School of Linguistics, Speech and Communication Sciences, Trinity College Dublin.

[11] Government of Ireland. 2022. Circular 0054/2022: Exemptions for the Study of Irish – Revising Circular 0052/2019. Retrieved from https://www.gov.ie/en/circular/28b2b-exemptions-from-the-study-of-irish-primary/

[12] Janette Dinishak. 2022. The deficit view and its critics. Disability Studies Quarterly, 36(4).

[13] Emer Smyth, Allison Dunne, Merike Darmody and Selina McCoy. 2007. Gearing up for the Exam: the Experiences of Junior Certificate Students. ESRI/DES. Dublin, Ireland.

[14] Ian Hardy and Stuart Woodcock. 2015. Inclusive education policies: Discourses of difference, diversity and deficit. International Journal of Inclusive Education, 19(2), 141-164.

[15] Peter-Paul Verbeek. 2006. Materializing morality: Design ethics and technological mediation. Science, Technology, & Human Values, 31(3), 361-380.

[16] Sara M. Acevedo and Emily A. Nusbaum. 2020. Autism, neurodiversity, and inclusive education. Oxford Research Encyclopedia of Education. Retrieved from https://doi.org/10.1093/acrefore/9780190264093.013.1260

[17] Breacadh. 2007. Liostaí Bhreacadh: Focail Choitianta sa Ghaeilge. Breacadh.

[18] Stephen Russell, Dewi Bryn Jones and Delyth Prys. 2022. BU-TTS: An Open-Source, Bilingual Welsh-English, Text-to-Speech Corpus. In Proceedings of the 4th Celtic Language Technology Workshop within LREC2022. Marseille, France, 104-109.

[19] Delyth Prys and Gareth Watkins. 2023. Language Report Welsh. In Rehm, G. & Way, A. (eds), European Language Equality: A Strategic Agenda for Digital Language Equality. Springer. 223-226.

[20] Bill Wake. 2003. INVEST in Good Stories, and SMART Tasks. Retrieved from http://xp123.com/ articles/invest-in-good-stories-and-smart-tasks/

[21] Naomi Winstone, Corinne Huntington, Lisa Goldsack, Elli Kyrou and Lynne Millward. 2014. Eliciting rich dialogue through the use of activity-oriented interviews: Exploring self-identity in autistic young people. Childhood, 21(2), 190-206.

[22] Eija Sevón, Marleena Mustola, Anna Siippainen, and Janniina Vlasov. 2023. Participatory Research Methods with Young Children: A Systematic Literature Review. Educational Review, 1–19.

[23] Cathy Engstrom and Vincent Tinto. 2008. Access without support is not opportunity. Change: The magazine of higher learning, 40(1), 46-50.

Developing New Bilingual Synthetic Voices for Children and Young People in Wales

A collaboration between NHS Wales, CereProc, and Bangor University

MEINIR WILLIAMS

Bangor University, Wales

DEWI BRYN JONES

Bangor University, Wales

SARAH COOPER

Bangor University, Wales

STEFANO GHAZZALI

Bangor University, Wales

DELYTH PRYS

Bangor University, Wales

Eight pairs of bilingual Welsh and English synthetic voices were developed for use by children and young people in Wales with speech difficulties who use Augmentative and Alternative Communication (AAC). These voices represent northern and southern Welsh accents in both English and Welsh. Four of the voices are female, and four are male. CereProc led the project, contributing their extensive expertise to build the voices. A commercial company was used to find and record the voice talents, and the Bangor University Welsh speech technology library was used to fine tune the voices and develop new components where needed. Feedback on the north Welsh voices was gathered by speech and language therapists, with preliminary results presented here.

Keywords: Welsh, Text to Speech, Synthetic Voices, Augmentative and Alternative Communication (AAC)

1 INTRODUCTION

Ever since the early days of personal computers, the idea of creating an artificial computer voice has been attractive. There are a number of reasons for this, including enabling visually impaired users to read text, and enabling non-verbal people to communicate orally. Text to speech technology was first developed in the 1960s for the English language, and as recent as 1990, they were almost exclusively male voices. In that year, in a notable development for gender equality, Ann Syrdal developed the first female voice [1]. The idea of having such technology working for Welsh at that time seemed far away, but in 2003 Bangor University in Wales and Trinity College, DCU and UCD in Dublin, Ireland obtained Interreg European Union funding to develop Welsh and Irish speech processing resources (WISPR) in a collaborative venture. This was a key development which lay down foundations for further research in speech and language technologies for both languages. The first voices developed were diphone based, and sounded rather robotic. They were however intelligible and were well received, especially by visually impaired users [2].

Alongside the text to speech research work also began on developing speech recognition for Welsh [3]. This technology encompasses both human speech directed at a digital device with the device responding to what was said, e.g. by turning on a light or searching for information, and also the transformation of the spoken word into

text, e.g. for transcription or subtitling. During this period, technology's ability to create better synthetic voices and better speech recognition was rapidly improving, especially following the adoption of AI and language model methods. Bangor University's Language Technologies Unit (LTU) developed a philosophy of publishing computer code and resources derived from their research under permissive open source licences, on platforms such as GitHub and the European Language Grid (ELG), with the Welsh National Language Technologies Portal serving as a one stop shop to obtain information on them all [4]. The intention was to enable commercial companies, as well as other researchers, to use the resources and further develop them, conscious of the fact that small languages like Welsh would not be included in speech technology applications unless doing so was easy and cheap to implement. Helped by funding from the Welsh Government, the Royal National Institute of Blind People (RNIB) commissioned Ivona, a company based in Poland, to create good quality synthetic voices to enable Welsh text to be read aloud by synthetic voices in order to aid visually-impaired people. This project resulted in a female voice (Gwyneth) and a male voice (Geraint) both of which were of higher quality than had previously been available for Welsh [5].

As the technology improved, it became possible to build synthetic voices that were a closer match to an individual's own voice. Companies and organisations began to offer synthetic voices to individuals who were about to lose their speech due to medical conditions that replicated their own voices. In Britain, this service was only available in English. One organisation who took the lead was the Motor Neurone Disease (MND) Association [6] who offered a voice banking service but did not have the expertise to deliver it for Welsh. With funding from the Welsh Government and support from the NHS and a number of speech and language therapists, the Lleisiwr project was created to provide a similar service for Welsh [7]. Because of the need to protect patients' and service users confidentiality, and guard against the risk of misuse of individuals' voices, these voices were not released publicly. In the meantime, it became obvious that synthetic Welsh speaking voices were needed for children and young people using digital communication devices, as it was not ideal for children and young people to have to use adult voices to communicate orally. The experience of developing Lleisiwr also showed that bilingual people in Wales regularly switch back and forth between Welsh and English, and it would be strange if an individual used different voices to speak in the two languages. This led the team to begin research into bilingual models that could enable a voice to switch easily between languages as needed, providing genuinely bilingual voices and pairs of voices.

Augmentative and Alternative Communication (AAC) systems have also come a long way, together with the need for improved voice quality. AAC includes any communication methods that help or take the place of oral speech. This can include a wide range of communication methods, from gestures and sign language to electronic devices that produce speech, with the various methods used depending on the individual's needs [8]. Prior to this project, individuals could use the Gwyneth voice with a basic Welsh language system on Smartbox Grid 3 software.

2 PROJECT DETAILS

In September 2021 NHS Wales announced a tender for the creation of Welsh language and Welsh-accented English language children and young people's voices to be used in AAC electronic communication devices. In response CereProc, a world-leading speech technology company based in Edinburgh, and the LTU at Bangor University, Wales, collaborated to present a joint tender. As a university based research centre, the LTU's emphasis is on research, avoiding competition with the private sector except in the case of market failure, and strives to collaborate with industry for the benefit of society and the economy. This resulted in a very happy partnership, with CereProc contributing their technological and commercial expertise, and Bangor contributing their expertise on Welsh language speech technology and linguistic and phonological matters. Bangor University had already released a text

to speech corpus of Welsh language adult voices [9], and had trained a synthetic voice which was available on the web under an open source licence [10]. Working with CereProc to produce high quality voices for NHS Wales seemed a natural progression from this. The project started in January 2023 and the voices were completed by the end of September in the same year.

Eight pairs of synthetic voices in Welsh and English for children and young people were developed with two geographical accents, for the north and south of Wales. This was felt necessary as there is no one standard non-dialectical pronunciation in Welsh and providing the users with voices that were as close as possible to the accent of their peers was one of the project's aims. Each language pair was voiced by the same voice talent in order to ensure that users could keep the same voice when speaking Welsh and English – an important consideration in a bilingual country. The voices are not themselves bilingual, but they do allow for some code switching. This means that individuals can use both languages in a much more integrated manner rather than if they were kept separate.

Voices needed to be created for a range of different ages in order to reflect the developing voices of children and young people. The commission asked for 4 voices for children aged 8-12, that would also be suitable for younger children, and 4 voices for teenagers aged between 13 -16 and upwards, that would be closer to adult voices. Details of the voices may be seen in Table 1.

Table 1: Voice details

Name	Gender	Accent	Age
Ffion	Female	Northern	13-16+
Seren	Female	Northern	8-12
Tomos	Male	Northern	13-16+
Owain	Male	Northern	8-12
Rhian	Female	Southern	13-16+
Catrin	Female	Southern	8-12
Rhodri	Male	Southern	13-16+
Gethin	Male	Southern	8-12

In order to create the voices, voice talents needed to be found who would reflect the age groups and geographical accents in both Welsh and English. The Darlun television production company were hired to undertake this task and auditioned over 300 children of different ages before sending samples of 16 voices to the team at Bangor and CereProc to make the final choice. The voice talents were recorded in studio reading aloud a script of 200 sentences designed to include all the phonemes of both English and Welsh, based on the Bangor Pronunciation Dictionary [11], adapted for this project, and a CereProc script which already existed for the English language voices. The Welsh and English scripts had about 1,200 sentences each for recording the voices, including sections which reflected the northern and southern accents. Sufficient examples of every phoneme had to be included in the scripts in order to build the voices, and extracts from children's novels, without breaching copyrights rules, were used to ensure that the material was of a suitable level to be read by the voice talent. Once the recording was completed, CereProc built the first versions of the voices.

Following this the process of normalizing the voices was undertaken. Since CereProc had already built a number of English voices the process of creating the English voices was fairly simple. However, a normalizer had to be developed to express numbers, units, times etc. for Welsh. A basic normalizer previously developed by Bangor

University for Welsh [12] was used as a basis. A number of problems had to be resolved concerning using the decimal and vigesimal systems in different contexts. The decimal counting system (11 = un deg un, 12 = un deg dau, etc.) was deemed more suitable for general use by children, following current school practice, while the vigesimal system is kept for dates e.g. 15th = y pymthegfed, and telling the time, e.g. twenty past three = ugain munud wedi tri, again in keeping with current best practice. The pronunciation of acronyms and abbreviations in Welsh is complex, with some using Welsh letter by letter pronunciation, e.g. S4C, some pronounced as one word, e.g. CBAC, and some pronounced according to the rules for English letters, e.g. BBC.

The issue of code switching between Welsh and English was also looked at as that is a very common phenomenon in current Welsh [13]. The English language voices also needed to be able to deal with some Welsh words, especially placenames. In order to achieve this, CereProc's English pronunciation dictionary was adapted to include Welsh phonemes, and rules for the pronunciation of Welsh were introduced to the English voices and vice versa. However, in instances where orthography conventions for the two languages clash, e.g. *Allan* in English /ælən/ and *allan* in Welsh /ałan/, the main language was kept for each language in order to ensure consistency.

The voices are available for download on Windows and MacOS systems and iOS and Android CereProc Voices apps. They will be suitable for use with a range of AAC software, including Smartbox, Liberator, Tobii Dynavox and Jabbla. It is hoped that further work will in future allow the voices to be available to anyone who would benefit from their use.

3 RESULTS

Dissemination is ongoing, as is the evaluation of the eight pairs of voices, especially in respect of ensuring that the voices are suitable for use in the contexts for which they were intended.

Beta versions of the Welsh and English northern voices were presented to a group of speech and language therapists, researchers and students interested in speech and language during the North Wales Speech and Language Exchange (NWSLE) during October 2023.

Thirty three participants filled in a follow-on questionnaire after the session. Seventeen were Speech and Language therapists, one was a Health Assistant and fifteen were either studying Speech and Language Therapy, were research students, or were assisting Speech and Language therapists.

Of these, twenty six were female, two male, three non-binary and two preferred not to say. Seven of the participants were aged between 18 and 24, seven between 25 and 34, fourteen between 35 and 44, and four between 45 and 54. Thirteen spoke Welsh fluently, seven spoke quite a lot of Welsh, five spoke a little Welsh and seven could say a few words. Sixteen noted that they worked in North west Wales, eight in mid north Wales, six in the north east, and two worked outside of north Wales.

During the session, there was a discussion of AAC in the local Health board (Betsi Cadwaladr) and the Assistive Technology Service (EAT) and a case study of AAC technology provided through the medium of Welsh, followed by a demonstration of the voices.

Samples of the voices speaking English (including Welsh placenames), formal Welsh including numbers, and informal Welsh including code switching and non-standard orthography were presented, as can be seen in Table 2.

Table 2: Samples of voices presented to attendees at the North Wales Speech and Language Exchange

Voice	English Sample	Welsh Sample	Welsh informal / code switching sample
Ffion	Hello, my name is Ffion and I come from Pwllheli. I am a digital voice for AAC	Ni ydi'r lleisiau digidol sydd wedi cael eu datblygu gan CereProc a Phrifysgol Bangor	Tisio panad?
Seren	Hello, my name is Seren and I come from Porthmadog.	Mae 4 llais o'r gogledd a 4 llais o'r de	Dani di bod yn Chester Zoo
Tomos	Hello, my name is Tomos and I come from Harlech.	Dwi'n un o'r lleisiau ar gyfer pobl ifanc 13-16 oed	O'dd o'n briliant
Owain	Hello, my name is Owain and I come from Wrexham.	A dwi'n un o'r lleisiau ar gyfer plant 8-12 oed	Ond dwi isio mynd adra

Following this, the participants were asked to complete a questionnaire giving feedback on the voices and the session more generally. The reactions of the attendees to the voices are given below. They were asked to respond to five statements and to note to what extent they agreed or disagreed with them on a scale of 1 (strongly disagree) to 5 (strongly agree):

- The voices sound natural
- The voices represent the north Wales accent well
- The voices represent differences between male and female voices well
- The voices represent the different age groups well
- The voices will represent children who use AAC well

Results can be seen in Figure 1. The attendees were also asked to answer a question about future developments in the field and what more they thought would be needed.

In general, the reaction was very favourable, with the majority of the replies either agreeing (4) or agreeing strongly (5) with the statements. Some comments were received stating that the accents did not reflect north east Wales as well, which is likely to be due to the fact that the voice talents for the north came from north west Wales. Also, it was noted that Ffion, the teenage female voice, sounded more like an adult, and one of the other responses noted that Tomos, the teenage male voice was high pitch, and therefore appeared younger.

A number of responders noted that further development of Welsh AAC software was needed to integrate with the voices, and that further development of dialectical variation in vocabulary, e.g. *llefrith/llaeth* (for *milk*) would also be welcome.

The mother of a 6 year old girl who had begun to use one of the Welsh Language voices also gave her feedback. The reaction to the voice itself was positive, and the mother emphasised how important it was to have a suitable voice for her daughter's age, saying "I'm so thrilled her voice is age appropriate now", but she also emphasised that the voices existed in a wider AAC system and that Welsh-medium provision should be developed comprehensively to enable individuals to communicate easily in Welsh and use the voices as widely as possible.

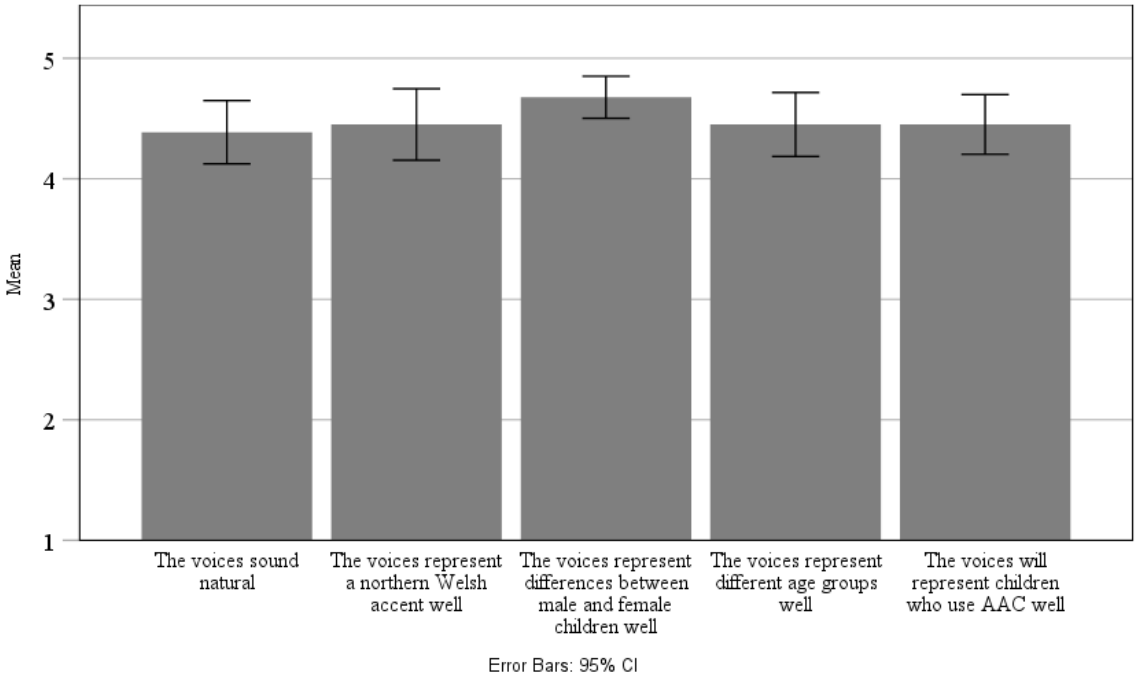


Figure 1: Participant questionnaire responses

4 CONCLUSIONS

The aim of this project was to create eight pairs of text to speech voices in both Welsh and English that would be suitable for use with AAC systems within a timetable of 9 months, and that aim was achieved. Creating pairs of voices in a minoritized language (Welsh) and a majority language (English) using the same voice talent and corresponding scripts allow the possibility of using bilingual AAC without changing voices. It will be interesting to see whether other minoritized languages will follow this methodology and provide similar bilingual provision.

Initial response to the beta version has on the whole been very favourable, and releasing the voices will result in wider feedback by users from every part of Wales. This will help us measure which features have satisfied users' needs and what remains yet to be done.

Perhaps the main need for further development is the software within which the voices are embedded. More Welsh language AAC communication boards would enable users to adapt the voices to reflect their communication desires including their accents, vocabulary, and their ways of using code switching. Whilst the Welsh and Welsh-accented English voices are an important step in the right direction, their full potential is not yet realised. Further software incorporating the voices would ensure equal opportunities for individuals who wish to live their lives through the medium of both languages. An innovative project developing AAC aids for Irish[1] shows how Welsh could also develop, especially given the common challenges facing the Celtic languages, including grammatical mutations, word order, yes/no answers and prepositions.

[1] See previous chapter in this book.

This project has shown that it is possible to develop high quality resources for AAC through the medium of two languages, and is a possible solution for small languages existing in the shadow of larger ones in bilingual societies. It is hoped that this pattern lays the groundwork for developing resources to ensure a wider range of languages are made available to all who wish to use them.

ACKNOWLEDGEMENTS

We wish to give our heartfelt thanks to Cereproc, Darlun, the voice talents, the EAT Service of NHS Wales, Jeff Morris, NWSLE, Rebecca Day and everyone else who has supported the project.

REFERENCES

[1] Cade Metz. 2020. Ann Syrdal, Who Helped Give Computers a Female Voice, Dies at 74. New York Times (20.08.2020). Retrieved from https://www.nytimes.com/2020/08/20/technology/ann-syrdal-who-helped-give-computers-a-female-voice-dies-at-74.html

[2] Delyth Prys, Briony Williams et al. 2004. WISPR: Speech Processing Resources for Welsh and Irish. Pre-Conference Workshop on First Steps for Language Documentation of Minority Languages, LREC Conference, Lisbon, Portugal.

[3] Sarah Cooper, Dewi Bryn Jones and Delyth Prys. 2014. Developing further speech recognition resources for Welsh. In Judge, J., Lynn, T., Ward, M. and Ó Raghallaigh, B. (eds), Proceedings of the First Celtic Language Technology Workshop at the 25th International Conference on Computational Linguistics (COLING 2014). Dublin, Ireland, 55-59.

[4] Delyth Prys and Dewi Bryn Jones. 2018. National Language Technologies Portals for LRLs: a Case Study. Lecture Notes in Artificial Intelligence. Springer.

[5] RNIB Cymru. Lleisiau synthetig Cymru. Retrieved from https://www.rnib.org.uk/cy/nations/cymruwales/lleisiau-synthetig-cymru/

[6] Motor Neuron Disease Association. 2022. Voice Banking for Motor Neurone Disease. Retrieved from https://www.mndassociation.org/sites/default/files/2023-05/P10%20Voice%20banking%202022%20v2.pdf

[7] Bangor University's Language Technologies Unit. 2018. Lleisiwr. Retrieved from https://lleisiwr.techiaith.cymru/

[8] Kristi L. Morin, Jennifer B. Ganz, Emily V. Gregori, Margaret J. Foster, Stephanie L. Gerow, Derya Genç-Tosun and Ee Rea Hong. 2018. A systematic quality review of high-tech AAC interventions as an evidence-based practice. Augmentative and Alternative Communication, 34(2), 104–117.

[9] Bangor University's Language Technologies Unit. 2021. Corpws Talentau Llais. Retrieved from https://git.techiaith.bangor.ac.uk/data-porth-technolegau-iaith/corpws-talentau-llais

[10] Stephen Russell, Dewi Bryn Jones and Delyth Prys. 2022. BU-TTS: An Open-Source, Bilingual Welsh-English, Text-to-Speech Corpus. In Proceedings of the 4th Celtic Language Technology Workshop within LREC2022. Marseille, France, 104-109.

[11] Bangor University's Language Technologies Unit. 2021. Geiriadur Ynganu Bangor | Bangor Pronouncing Dictionary. Retrieved from https://github.com/techiaith/geiriadur-ynganu-bangor/tree/21.03

[12] Bangor University's Language Technologies Unit. 2023. techiaith-tts. Retrieved from https://github.com/techiaith/techiaith-tts/tree/main

[13] Margaret Deuchar and Peredur Davies. 2009. Code switching and the future of the Welsh language. International Journal of the Sociology of Language, 195 (Jan. 2009), 15-38.

Addendum

MÉLANIE JOUITTEAU - L'UNIVERSITÉ BORDEAUX MONTAIGNE AND L'UNIVERSITÉ DE PAU ET DES PAYS DE L'ADOUR

AI is coming and our languages need to catch the wave

My point is simple. Our languages must become AI compatible if they are to survive the coming deployment of applications based on personal data.

Artificial Intelligence (AI), a system characterized by models that are trained using extensive datasets, including linguistic data, has become an integral part of our daily lives. This is evident in our frequent interactions with predictive typing and voice recognition technologies on smartphones. Currently, we are witnessing only the initial phase of its societal integration. With this context in mind, I draw a parallel between the social integration of AI and that of mobile phones. In a detailed analysis comprising twelve points, I outline the manner in which the previous technological revolution, namely mobile phones, became embedded in our lifestyles. I draw comparisons with the ongoing AI revolution at each of these stages.

The present phases in AI development are laying the groundwork for the implementation of applications reliant on personal data. I offer a detailed explanation of why, from my perspective, AI applications that utilize personal data pose a significant risk of drastically reducing the world's linguistic diversity in the near future. Although the swift integration of AI into our lives seems almost inevitable, the specific threat of linguistic homogenization can be mitigated. This mitigation is possible through the adaptation of our languages to be compatible with these AI models.

I advocate for a collective effort encompassing public, industrial, scientific, and societal sectors. This mobilization is crucial to ensure that human languages successfully navigate, rather than drown in, the AI tide.

Do you possess a mobile phone?

As a member of Generation X, I witnessed the advent of mobile phones and smartphones. This experience has led me to draw parallels between societal responses to mobile phones 25 years ago and contemporary reactions to AI. It appears to me that society's response to these technological revolutions bears striking similarities. In the 1990s, individuals who were initially hesitant about mobile phones now routinely carry them. I have methodically catalogued the various factors that contributed to this shift in attitude, tracing the technology's societal penetration from its nascent stage to complete integration and points of no return.

Let us reflect on this transition. Those who experienced this shift, particularly those who were initially resistant to mobile phones, can each recall their personal moment of capitulation. Similarly, observers of AI's evolution can discern parallels with the current stage of its development.

In the beginning, it's always…

1 THE AGE OF THE PRIVILEGED FEW

This phase is characterized by the rapid adoption of new technology by a small group of enthusiasts who possess significant economic, cultural, and technological power. Such adoption bestows upon the technology a symbolic status of high social standing. The 'Age of the Privileged Few' continues if the technology remains inaccessible to the majority. In contrast, if the technology becomes widely affordable, it is then described as 'democratizing.' This term is misleading because it bears no democratic value; it merely indicates that the product is no longer exclusive to an elite group. This process does not involve democratic procedures like national elections or referendums to determine the adoption of new technology. Market forces predominantly drive these changes. This phase was unmistakably evident in the case of mobile phones and, since the advent of ChatGPT, has increasingly applied to AI. Notably, ChatGPT offers some of its services for free, unlike mobile phones, further accelerating the spread.

2 PROFESSIONAL REQUIREMENTS

The advancement of technology alleviates numerous professional duties and becomes deeply ingrained in various occupations. The ubiquity of mobile phones has facilitated constant connectivity for work-related tasks, a development that was initially welcomed and eventually became a necessity across many professions, particularly in fields like healthcare where doctors on call must be readily accessible. Especially in an era characterized by widespread unemployment, resistance to adopting mobile technology increasingly results in an exit from these professions. AI is increasingly providing a competitive edge in numerous sectors, including programming, content creation, and customer services. The efficiency of AI is also paving the way for a potential surge in unemployment, providing both means and context for an acceleration of this effect.

3 REDUCING MAJOR MALFUNCTIONS

The common perception of mobile phones as 'dangerous and ineffective' prevailed in their early days. Initially, their portability was a subject of ridicule, given their limited signal reception, which was largely confined to small areas within major urban centres. Despite persistent and significant concerns regarding the health risks posed by the radiation from these early models, the user base and network coverage have expanded considerably. However, it is crucial to note that two primary issues from the inception of mobile phones persist:

1. individuals sensitive to electromagnetic waves continue to seek refuge in the few remaining 'white zones', often with little societal concern for their wellbeing, and
2. cellular reception remains inconsistent across various locations.

Despite these ongoing challenges, mobile phones are widely used. They are 'good enough'.

Similarly, as of 2024, AI technology, including platforms like ChatGPT, occasionally generates erroneous responses, and it remains uncertain whether these 'hallucinations' can be entirely eliminated. AI developers know that delivering flawless technology is not a prerequisite for widespread adoption; rather, the goal is to achieve a level of functionality that meets the threshold of 'good enough' for most users. It is already 'good enough' for millions of users who feed interfaces like chatGPT with masses of personal and industrial data and provide human feedback on the results, for the technology to improve, and for capital to flow in to reduce hallucinations.

4 FEARS AND EMERGENCIES

Fear serves as a potent catalyst for social change. The use of emergency communication systems and the nomadic accessibility of crucial information often overcome entrenched hesitations. In emergency situations, the potential advantages, such as immediate access to assistance or essential information, often outweigh other considerations. This is particularly true when it concerns the safety of oneself or loved ones, especially for the most vulnerable among us.

It is evident that AI applications will be regarded as beneficial to our health and safety, akin to how smartwatches monitoring vital signs are perceived. This is especially relevant for children, the elderly, and individuals with disabilities. Notably, these populations are frequently at the forefront of technological integration trials involving body-close technologies, such as implants. These trials may pragmatically challenge deep-seated societal taboos regarding the physical integration of technologies.

5 ENHANCING HUMAN CONTACT THROUGH TECHNOLOGY

The innate human penchant for interpersonal communication has been significantly augmented by emerging technologies. Smartphones, for instance, have added a layer of immediacy to our interactions with family and friends through instant messaging and social networking. This advancement had promised enhanced human contact, facilitating communication at a rapid pace and consistency that traditional landline telephones could not sustain. The phenomenon of FOMO (Fear of Missing Out) has further amplified this trend. Among younger generations, it manifested as anxiety over missing social gatherings or updates within their social circles. For their parents, it often reflected a concern about losing connection with their children. In the era preceding the widespread adoption of the internet, marked by mass unemployment, this urgency extended to adults anxiously awaiting job opportunities via telephone calls – a scenario that preceded the ubiquity of dating apps.

Regarding AI, its role will not be to create distance from our loved ones. Instead, AI will present itself as a better method of connecting with people. AI-powered chatbots generate human-like language on demand, tirelessly and without restrictions. They have the capacity to produce language that is quintessentially human, at a volume and with a consistency unattainable by any actual human being.

The appeal of these technologies lies in their promise of increased human interaction. This very characteristic underpins their capacity to collect extensive personal data, as we frequently utilize these technologies in our most private relationships. It is evident that citizens are progressively becoming more educated about safeguarding their personal data. Some argue that the advancements in technology integration simultaneously foster social resistance towards these technologies. However, I believe this perspective confuses resistance to the implementation of technologies with the transformation of our human cultures in response to these technologies. Rather than manifesting as resistance, this cultural adaptation facilitates their harmonious integration. It fundamentally lays the groundwork for the viability of applications reliant on personal data.

At the next stage, starting with point 6, we will want to offer AI applications our most intimate and personal data. This is when the speakers of languages that are not compatible with AI will begin to be penalised.

6 CONVENIENCE AND EFFICIENCY

The statement, "I don't like this stuff, but I dislike paperwork even more," reflects a sentiment shared by many. A common admission is that the desire for convenience has driven the transition to mobile phones, as these devices enable the efficient handling of mundane tasks. Smartphones indeed offer an extensive array of applications and services that enhance daily life.

These include spatial navigation through GPS, time management via integrated calendars, and simplified financial transactions through mobile banking.

AI applications, designed to utilize personal data, aim to satisfy the demand for cohesive management of numerous monotonous tasks. In the healthcare sector, immediate access to personal data enables the comprehensive integration and real-time analysis of various health parameters. These include bodily metrics, dietary habits, stress levels (quantified and analysed), genetic makeup, and even environmental conditions such as the weather. This integration facilitates the generation of preliminary diagnoses, which can then be further examined by medical professionals' AI systems. The efficiency and potential benefits offered by such technological advancements are indeed noteworthy.

7 CENTRALISATION OF TOOLS

The integration of diverse services and functions into smartphones, including alarm clocks, cameras, email, messaging, social media, and even tide tables, have created the perception of liberation from a clutter of individual devices, all potentially replaceable by a single mobile phone.

AI applications, utilizing personal data, aim to personalize experiences by adapting to individual interests, behaviours, and preferences, thereby managing the plethora of apps in a manner that suits your unique style and pace. Ultimately, this seeks to liberate users from the overwhelming array of apps on their mobile devices. It seems unlikely that we will resist this trend.

8 ACCESS TO INFORMATION

The integration of the internet with smartphones has made a vast repository of information readily accessible, serving as a significant motivational factor. Individuals who were initially sceptical about mobile phones likely compared their concerns with the advantages of having immediate access to news, knowledge, and various resources.

Large corporations and governments implement information filtering mechanisms, utilizing algorithms to pre-select the information available through various channels, such as the filters currently employed by Facebook. AI, when supplied with personal data, has the power to revolutionize this filtering process. AI promises enhanced individual control, enabling users to tailor their internet experience based on trusted sources and personal interests, rather than relying on the predefined filters of others, including politicians and advertisers. Note that the upstream filtering will probably persist, but the individuals will be able to subtract further. For instance, individuals with a phobia of snakes can avoid encountering such images online. Those concerned with cultural taboos not widely acknowledged by society can navigate the internet more comfortably. Their digital assistant will exclude content on demand. Thus, each person's access to the world of information can become more personalized, shaped by their own preferences.

The next stage starts with point 9. This stage is characterised by a loss of control, points of no return and restrictions on individual choices.

9 DATA SOVEREIGNTY, SECURITY

In recent years, smartphones have evolved to include biometric authentication and advanced encryption technologies. Notably, enhancements in privacy settings, application permissions, and user control over data access have gained prominence. Initially,

privacy concerns led to hesitancy among users; however, these advancements have bolstered confidence in safeguarding personal information. Observe that this increased confidence persists despite ongoing issues, such as the persistent tracking of users' geolocation. Individuals travel daily or internationally, carrying sensitive information ranging from banking details to very intimate data in their devices, perceived as secure 'digital vaults'. Losing a smartphone is not uncommon. Certain customs authorities also legally have the right to fully access this data. This is where our situation is increasingly analogous to the paradigm of addiction, where the desire for technology overrides critical security considerations. Our security concerns are often allayed by superficial reassurances once we need access to it (or once we think we do). Consequently, we tend to neglect fundamental rules of personal security and overlook significant vulnerabilities that could impinge on our basic freedoms.

10 SOCIAL INTEGRATION

With the widespread adoption of smartphones, social pressure for their use has intensified. Friends, family, and colleagues often depend extensively on smartphones for communication and coordination. This makes it challenging for individuals to refrain from using these devices without experiencing feelings of isolation or exclusion. At this stage, technology becomes an integral part of human relationships. It not only extends these relationships by facilitating them. It is part of them, and plays a crucial role in shaping them. After such cultural mutation, human interactions deprived of technology are impaired and incomplete, lacking the support of the cultural framework that existed prior to the advent of technology.

11 TOTAL UBIQUITY OF SERVICES

A wide range of essential services, including transportation, retail shopping, ticketing, and banking, have become increasingly integrated with smartphone applications. Following the transition to digital platforms, service providers often find it costly to accommodate consumers who do not comply with these digital norms. Consequently, the access rights of these non-compliant consumers to services gradually come into question, leading to a widening of social divides. At this stage, and thinking back to the example we took with health applications, general practitioners may no longer receive training to treat patients whose personal data has not been pre-processed by AI.

12 CULTURAL AND COGNITIVE POINTS OF NO RETURN

Social practices have evolved to incorporate the use of smartphones. It is now common, and often considered polite, to confirm appointments en route, to alter the meeting location at the last minute based on preference, or to notify others of even minor delays. Such practices, integral to modern social rituals, are unfeasible without this technological aid. Generations who have witnessed this transformation recall a time when it was possible to schedule appointments months in advance and a common courtesy to arrive at the designated time and place without any need for further confirmation. Additionally, our collective ability to memorize telephone numbers has diminished as we increasingly rely on smartphones for this task. These changes represent cognitive points of no return. Organized resistance to this trend would necessitate the ability to coordinate without such tools, a skill progressively fading. The only option for individuals who resist this dependence is to reject the technology, albeit at the cost of social isolation.

CONCLUSION

This article is not a prophecy but rather a prospective analysis. In this narrative of our integration into the mobile technology revolution, is there truly a fundamental difference, either contextually or technologically, that distinguishes it from AI? From my perspective at the beginning of 2024, I believe we have moved largely beyond phase 3. Major industries are evidently

preparing for subsequent developments, and I perceive no barriers to this ongoing evolution of technology integration, reminiscent of this other technological revolution I have witnessed in my adult life. If, like me, you discern no fundamental disparity between the societal adoption of these two technologies—mobile phones and AI—it seems logical to infer that resistance to AI is unlikely. The influence of this wave will rapidly extend to our most personal spheres, likely impacting even our physical bodies and cognitive processes. Any individual or collective resistance will amount to mere cultural adaptations to a new reality. This evolution presents and will continue to present multiple challenges, which the scientific community is currently ill-equipped to address [1]. AI technologies demand substantial hardware capabilities for processing vast amounts of data. A limited number of competing companies are making significant decisions that affect our lives and our future as a species. These decisions are made without democratic input and without the necessary internal resources to fully consider the wide-ranging societal implications of their actions. These companies do not employ scientists from the humanities to begin to address theses problems.

As a linguist, it is the danger of a drastic reduction in linguistic diversity that appears particularly salient to me. To illustrate with a concrete case, the French state has a single official language, French, which will undoubtedly be incorporated into AI technologies. The more than one hundred other unofficial, unprotected languages of the citizens of the French state are not at all equipped to overcome the AI barrier. Most of them are in a state of digital insecurity, which we are still struggling to map [2].

There is a widespread belief in our societies, apart from within a few influential but isolated circles, that resistance to artificial intelligence will be strong in society, and that forecasts of technological revolution have erred by not taking sufficient account of the resistance it will trigger. In Brittany, I come across this belief in very different generations and very different social circles, from the Monts d'Arrée to the high-tech academic circles of the megacities, as well as in... the high-tech circles of the Monts d'Arrée. I respect this belief as such. As a belief. I respect the emotions it protects, and the people who feel them. But I think it is objectively inaccurate. I also think that this belief is dangerous when it slows down the measures we can still take to adapt to this coming mutation. This change is objectively inevitable. The speakers need to be aware of the particular dangers that the development of AI will pose to their linguistic practices if the new tools cannot work in their languages. Denying it slows down the implementation of emergency measures to counter these dangers. If our languages cannot be processed automatically, so as to integrate the new tools, they will face an even more radical reduction in practice than they did in the twentieth century. Most of them will disappear.

Anyone who knows anything about the history of minority languages in the twentieth century should be on high alert right now. Here's why. We know that languages disappear when parents think that these languages are dangerous for their children. Non-AI compatible languages will become dangerous for their users, as the integration of the new apps progress in society. We have seen that there will be AI applications to which we - ourselves - will want to offer our intimate and personal data. These are points 6 to 8 illustrated above. If small languages cause these applications to malfunction, we will replace them with the languages that let them function properly. This is already the case in practice when we simply dictate an SMS in French to take advantage of its automatic writing rather than typing it in Breton. But the scale of avoidance is going to be immense, unparalleled. When society reaches the points of no return, the points between 9 and 12 of total and irreversible integration, the pressure to eliminate malfunction sources will increase. Nonconformity will come with social exclusion, because these tools settle at the heart of human relationships. When these applications become essential for access to banking, justice, education, dating and health services, we will seek to protect our children's future by reluctantly forbidding them to use our non-AI compatible

languages, which will be objectively transformed into a serious social handicap, in a great global movement towards monolingualism. History will say that we consented to this.

There are still between 6,000 and 7,000 languages in the world. They are the products of our human bodies, of billions of human brains interacting with each other over thousands of generations. Each disappearance of a language produces deep wounds, and echoes over several generations. Imagine being raised by people who refuse to speak their own native language to you. And we are very bad at healing those wounds. Our generations have the opportunity to mobilise to prevent this harm to be done, to ensure that this diversity is not wiped out, and that AI is an opportunity for multilingualism and the preservation of this incredible diversity. It undoubtedly has the power to be for languages made compatible with this technology.

We need a global rescue plan, a lively public, scientific and societal mobilisation, a collective rescue plan for our linguistic treasures. In practice, languages need a digital data set of written and spoken language, put into the right form for digital processing [3]. It is now crucial and urgent for our language communities to build up corpora of written and spoken language and make them accessible and open, to scientists and industry alike, with open copyright. You might think that the urgency derives from their actions in the first place, and you might also be very angry. However, preventing them from accessing the AI integration material for your languages is resistance of a purely moral nature, and is not effective in improving the situation. Resistance to AI is not going to happen. Taking languages hostage would seal their fate. Swimmers know that when the wave is too big, you have to dive in. Not just stand there and say that the wave shouldn't be there. Language support policies need to rethink their arsenal of support by organising them around the necessary surviving digital toolkit [4]. They need to encourage the emergence of trans-disciplinary collaboration cells between people who know how to code, to educate AI in languages with restricted corpora, and people who know how to talk to people, harvest language material in an ethical manner, and enrich it with precise standard annotations. Language communities need to be digitally empowered, in order to build the digital resources they need to be integrated into the new tools.

As for the scientific communities that produce, process and work on corpora, my very dear colleagues, descriptivist and formal linguists, field workers, corpus and elicitation linguists, annotators and engineers of data architecture, we have an historic opportunity to safeguard the object of our research for human societies.

It's time to rise to the occasion and show them what we can do.

ACKNOWLEDGEMENTS

I would like to thank Rayan Ziane, Loic Grobol, Reun Bideault and Milan Rezac for their critical reviews and discussions on the matter. Any misinterpretations and shortcomings remain my own.

REFERENCES

[1] Samuel R Bowman. 2023. Eight Things to Know about Large Language Models. Apr. 02, 2023. arXiv:2304.00612.

[2] Mélanie Jouitteau, Sylvain Kahane and Loic Grobol. 2023-present. Entrelangues, second edition, IKER & Modyco, CNRS. Retrieved from https://entrelangues.modyco.fr/

[3] Steven Abney and Steven Bird. 2010. The Human Language Project: Building a Universal Corpus of the World's Languages. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics ACL, Uppsala, Sweden, 88–97.

[4] Mélanie Jouitteau. 2023. Guide de survie des langues minorisées à l'heure de l'intelligence artificielle : Appel aux communautés parlantes. Lapurdum, numéro spécial 6. Retrieved from https://hal.science/hal-04090195v2