Language Report Welsh
Prys, Delyth; Watkins, Gareth

European Language Equality

DOI:
https://doi.org/10.1007/978-3-031-28819-7

Published: 07/06/2023

Publisher's PDF, also known as Version of record

Cyswllt i’r cyhoeddïad / Link to publication

Dyfyniad o’r fersiwn a gyhoeddwyd / Citation for published version (APA):
European Language Equality: A Strategic Agenda for Digital Language Equality (pp. 223-226).
(Cognitive Technologies). Springer. https://doi.org/10.1007/978-3-031-28819-7

Hawliau Cyffredinol / General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or
other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal
requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private
  study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to
the work immediately and investigate your claim.
Chapter 37
Language Report Welsh

Delyth Prys and Gareth Watkins

Abstract In this chapter, based on Prys et al. (2022), an update to the META-NET White Paper (Evas 2014), we present Language Technology (LT) for the Welsh language, providing an overview of the status of Welsh in Wales and a summary of the Welsh writing system and typology. We describe key tools and our recommendations for Welsh LT and associated resource development.

1 The Welsh Language

Welsh is mainly spoken in Wales, together with a small population in Argentina. A minoritised language (Prys 2006), Welsh is considered “vulnerable” (Moseley 2010). Welsh has official status in Wales (National Assembly for Wales 2011). The 2011 census reported that there were 562,000 Welsh speakers in Wales (19% of the population). The Welsh Government aim to almost double that figure by 2050 and recognise that technology is key to this ambition (Welsh Gov. 2017).

The Welsh alphabet contains 29 letters, including eight digraphs (e.g., ch) and the letter j borrowed from English to represent the borrowed /dʒ/ consonant phoneme. V, x and z are not used in Welsh, but are included with the alphabet for computer use as they often appear in named entities such as foreign placenames. Welsh belongs to the insular Celtic branch of Indo-European languages. It is verb initial, following a VSO order. It has consonant mutations at the beginning of words. Accented characters are common over vowels. Welsh has a continuum of other registers, with colloquial or informal registers differing markedly from the standard written form. It has many local dialects, with the main difference between those of north and south Wales. Welsh has two methods of verb formation, utilising concise forms or periphrastic forms, using auxiliary verbs. Guidelines to the latest version of the modern Welsh orthography, first standardised in 1928, were published in 1987 (Prys 2006). In 2021 a new Welsh Orthography Panel was established by the Welsh Government, which aims to resolve minor inconsistencies in the orthography.

Delyth Prys · Gareth Watkins
Bangor University, United Kingdom, d.prys@bangor.ac.uk, g.watkins@bangor.ac.uk

© The Author(s) 2023
2 Technologies and Resources for Welsh

According to Cunliffe et al. (2021), “on the Digital Language Vitality Scale […], Welsh is ‘Developing’, arguably tending towards ‘Vital’ in some aspects”. 90% of the 2019/2020 National Survey for Wales’ respondents used the internet (Welsh Gov. 2021). However, English is the dominant online language among Welsh speakers (Welsh Gov. 2015). A lack of language tools for Welsh and inequality or lack of equivalence to English language provision exacerbates the problem.

The major paper dictionaries have been digitised and made available online, and ongoing lexical work now occurs natively in a digital environment. In contrast to traditional descriptive dictionaries, terminology work in Welsh is concept based, held in databases, and published in many formats. These resources have been re-used in lexicons for various purposes, including spelling and grammar checkers.

Monolingual, bilingual and multilingual text corpora, as well as speech corpora, mainly in the standard or neutral language register, have been curated. The Language Technologies Unit at Bangor University holds the largest collection of corpora, at over 700 million tokens, including the Cysill Ar-lein Monitor Corpus (Prys et al. 2016). The CorCenCC (Knight et al. 2020) corpus is the largest annotated, balanced general corpus to date, with 11 million tokens. Crowdsourcing has been successfully used to gather large speech corpora of recorded prompts, currently using Mozilla Common Voice. Recordings of voice talents, collected specifically for building synthetic voices, have been released under the CC0 licence. Intellectual Property and licensing issues are of utmost concern when assessing the suitability of these corpora for use and reuse and can hamper their open distribution.

In terms of speech technology, a Welsh personal assistant (Jones 2020) has been developed as has the first Welsh speech-to-text transcriber. Synthetic voices have been created for Welsh using older diphone technology, with newer, more natural sounding unit selection voices becoming available under open licences. A voice banking initiative, Lleisiwr, a joint venture between Bangor University and NHS Wales, has been created for bilingual Welsh/English speakers about to lose their speech capabilities, and is one of the most innovative services established to date.

Acoustic and language models for Welsh are being developed. Some of these are part of multilingual sets, which are of variable quality compared to those developed specifically for Welsh. A Welsh part-of-speech tagging model has been developed for spaCy, unlocking the potential to perform many other NLP tasks on Welsh texts. Welsh has NLP tools for text analysis, anonymisation, and information extraction.

In terms of translation, a commercial Welsh–English translation system exists and MT for Welsh is offered by some major companies such as Google and Microsoft. Moses has been used to develop SMT for Welsh. Newer neural net engines are being used, and the first domain-specific MT engine for health launched. Welsh/English translation memories can be shared on the Open Translation Memories site, emulating the ELRI project. An overview of these LT tools and resources may be found on the Welsh National Language Technologies Portal (Prys and Jones 2018).

While the UK LT industry is mostly focused on the English language, Welsh language LT provision is mainly driven forward by the higher education sector. Wales
has vibrant creative technology, media and translation sectors which make use of
the government-funded open source LT created by universities. The main hub for
LT research in Wales is Bangor University, notably its Language Technologies Unit.
Relevant research is also undertaken at the universities of Cardiff, Swansea and
South Wales. Efforts have also been made to improve teaching digital technologies in
schools and universities. The current Welsh Government’s Welsh language strategy
states that “We must ensure that high-quality Welsh language technology becomes
available [...] to support education, workplaces and social use of Welsh” (Welsh
Gov. 2017). This was further elaborated in the Government’s Welsh Language Tech-
nology Action Plan (Welsh Gov. 2018). After years of small-scale and fragmented
initiatives, the publication of this plan provides a coherent, planned way forward for
the development of Welsh LT resources, tools and services.

3 Recommendations and Next Steps

There has been much progress in Welsh LT in recent years, but further work needs
to be done if the Welsh language is to thrive in the digital world. While FAQ gen-
eration is used for the Welsh language, the development of more sophisticated chat-
bot systems would further benefit Welsh speakers. There is no published research
on Welsh language knowledge graphs, nor what they have to offer to Welsh. Lim-
ited research has been conducted on Welsh language sentiment analysis. A key new
area for development is bilingual models to aid minoritised languages where users
constantly have to switch between their own language and the majority language
or code-switch within the minoritised language. Promising work has been done for
Welsh in developing a bilingual model for text-to-speech. Similar work for speech
recognition is underway, where pre-trained multilingual acoustic models can pro-
vide useful crosslingual speech representations that can be fine-tuned for effective
bilingual Welsh and English speech recognition. There are many other bilingual sit-
uations where a similar approach could be explored.

In order to fill these gaps Welsh needs to be able to join in large-scale multi-
national and multilingual research and development programmes of the type previ-
ously reserved for official EU languages. Also, in common with other minoritised
languages, Welsh needs a space within the European community where special atten-
tion can be paid to up-resourcing these languages and up-skilling their communities.
Minoritised European languages often also belong to the economic periphery in Eu-
rope, and using LT for economic regeneration in those areas would have a positive
effect on their economic, social and linguistic well-being.

It is often more attractive to court new and exciting project ideas. Funding oppor-
tunities are often prejudiced in favour of such ventures, but attention also needs to be
paid to maintaining, improving, consolidating and further developing existing tools
and resources. At the same time minoritised languages need to take full advantage of
any emerging innovations, playing their full part in the LT developments for Europe.
References


Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.