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

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_37

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 transciber. 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 Technology 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 generation is used for the Welsh language, the development of more sophisticated chatbot systems would further benefit Welsh speakers. There is no published research on Welsh language knowledge graphs, nor what they have to offer to Welsh. Limited 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 provide useful crosslingual speech representations that can be fine-tuned for effective bilingual Welsh and English speech recognition. There are many other bilingual situations where a similar approach could be explored.

In order to fill these gaps Welsh needs to be able to join in large-scale multinational and multilingual research and development programmes of the type previously reserved for official EU languages. Also, in common with other minoritised languages, Welsh needs a space within the European community where special attention can be paid to up-resourcing these languages and up-skilling their communities. Minoritised European languages often also belong to the economic periphery in Europe, 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 opportunities 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

Cunliffe, Daniel, Andreas Vlachidis, Daniel Williams, and Douglas Tudhope (2021). "Natural language processing for under-resourced languages: Developing a Welsh natural language toolkit". In: *Computer Speech & Language* 72.

Evas, Jeremy (2014). *Y Gymraeg yn yr Oes Ddigidol – The Welsh Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/welsh.

Jones, Dewi Bryn (2020). "Macsen: A Voice Assistant for Speakers of a Lesser Resourced Language". In: *Proceedings of the 1st SLTU-CCURL workshop*. Marseille, France: European Language Resources Association (ELRA).

Knight, Dawn, Steve Morris, Tess Fitzpatrick, Paul Rayson, Irena Spasić, and Enlli Môn Thomas (2020). *The National Corpus of Contemporary Welsh: Project Report; Y Corpws Cenedlaethol Cymraeg Cyfoes: Adroddiad y Prosiect*. https://corcencc.org/wp-content/uploads/2020/06/CorCenCC-report_2020_en.pdf.

Moseley, Christopher (2010). *Atlas of the World's Languages in Danger*. Paris: UNESCO.

National Assembly for Wales (2011). *Welsh Language (Wales) Measure 2011*. https://www.legislation.gov.uk/mwa/2011/1/contents/enacted.

Prys, Delyth (2006). "Setting the Standards: Ten Years of Welsh Terminology Work". In: *Terminology, Computing and Translation*. Ed. by Pius ten Hacken. Tübingen: Narr.

Prys, Delyth and Dewi Bryn Jones (2018). "National Language Technologies Portals for LRLs: A Case Study". In: *Human Language Technology. Challenges for Computer Science and Linguistics*. Cham: Springer.

Prys, Delyth, 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 LREC 2016*. Portorož, Slovenia: European Language Resources Association (ELRA).

Prys, Delyth, Gareth Watkins, and Stefano Ghazzali (2022). *Deliverable D1.34 Report on the Welsh Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-welsh.pdf.

Welsh Gov. (2015). *Welsh language use in Wales, 2013–15*. https://www.gov.wales/sites/default/files/statistics-and-research/2018-12/160301-welsh-language-use-in-wales-2013-15-en.pdf.

Welsh Gov. (2017). *Cymraeg 2050: A million Welsh speakers*. https://www.gov.wales/sites/default/files/publications/2018-12/cymraeg-2050-welsh-language-strategy.pdf.

Welsh Gov. (2018). *Welsh language technology action plan*. https://www.gov.wales/sites/default/files/publications/2018-12/welsh-language-technology-and-digital-media-action-plan.pdf.

Welsh Gov. (2021). *Internet skills and online public sector services (National Survey for Wales): April 2019 to March 2020*. https://www.gov.wales/internet-skills-and-online-public-sector-services-national-survey-wales-april-2019-march-2020-html.