PLANNING FOR LANGUAGE TECHNOLOGY DEVELOPMENT

Olimov Sardorbek Samandarovich
Trainee teacher of Namangan Engineering and Construction Institute

Abstract*: Welsh speakers have striven to maintain and revitalize their language in the face of the increasing dominance of English. Language technologies, the internet and digital media important to make Welsh more attractive, relevant and accessible.*

Key words*: language technology planning, best practice.*

The release of data itself under open source licences hasbeen the subject of some debate in Wales as elsewhere. Large datasets are one of the core requirements to train any models in speech technology, MT and conversational AI agent applications. Finding enough appropriately licenced data is one of the biggest challenges for less-resourced communities. Any strategy for efficient and effective harvesting of data can make the difference between supporting a language or not in a software package. For example, attempts have been made with apps specifically developed to crowdsource a speech corpus from the Welsh language community (Cooper, Jones and Prys, 2019). More recent activity in crowdsourcing Welsh language speech data has focused on collaborating with and sharing efforts with Mozilla's CommonVoice initiative, since its philosophy and motives align (Prys and Jones (2018 (1)). Not all language communities are happy to lose control of their data, especially where they have had bad experiences of colonial exploitation in the past.

However, in the Welsh context, the use of crowdsourcing strategies and of permissive licensing of data has helped the development of Welsh language software by the private sector, aided in some cases by knowledge transfer partnerships between academia and industry. Even in Wales it has not always been possible to release data on open source licence, as some legacy products came with their own, previous licences. In other cases, there was the need to sell commercial software in order to fund the continuation of the work. Increasingly however, and wherever public funding was used to create the tools and resources, they were released on permissive licences such as BSD, MIT, Apache or CC-0, which permit reuse without any restrictions. This was in order to make the tools and resources attractive to enable both small and large companies to take up and use in their own products. In both cases, the private sector is less willing to take up tools and resources published under copyleft licences, such as GPL and CC-BY-SA, which stipulate that the entire utilising body of software must be released openly under the same licence. Although Welsh, with its approximately half a million speakers is deemed to be a very small market for commercial companies, it is still large enough to support many small companies who could benefit from language technology tools and resources. These include translation companies, local media, software companies, web designers, and producers of educational games and language teaching materials.

In common with the experience of many other minoritized and endangered languages in peripheral regions, there are high proportions of Welsh speakers in the rural and remote north and west of Wales, areas that are impoverished with few opportunities for well-paid employment and therefore suffer from emigration of young, talented people. Providing appropriately licensed language resources to small, local companies in these areas can therefore help make them viable and help economic as well as linguistic revitalization of these areas. The arguments for releasing resources on permissive licences for large multinational companies are somewhat different. It can be argued that multinationals can well afford the development costs of including smaller languages amongst their multilingual offerings, and that paying for the necessary linguistic resources would be a great help to

those languages. However, in the absence of strong legislation requiring Welsh language provision, most multinationals only heed the economic argument, and if the cost of producing or procuring those resources is larger than the anticipated return on their outlay, they will not pay for their development. If, on the other hand, appropriate resources are available to them at no cost, they are then more willing to consider supporting that language amongst their offerings.

The minoritized or endangered language community benefits as many of their users already use those products every day in English, Spanish, French or whatever other dominant language they speak. Additional clarity would however be welcomed in understanding the legal ramifications of different licences as there are many legal grey areas. For example, if new language or acoustic models are trained from a specific corpus, does the licence of the original corpus carry over to the new models? Or when a new MT engine is trained on a certain dataset, how does that affect the licencing of the new product? This is especially problematic when there are different licences for the two languages in a bilingual corpus, especially derived from a translation memory where the copyright of the original language text was not originally made explicit. If tools and resources are to be shared outside individual projects and institutions, then dissemination is another issue that comes to the fore. International repositories such as Metashare, Github and Docker Hub have made it easier for developers to find resources in different languages, but for the non-expert user, and anyone interested in a specific language, the plethora of different repositories can be confusing. In addition to using the international repositories therefore, a Welsh National Language Technology Portal was established as a ‘one stop shop’ or ‘brochure site’ pointing at the different resources and giving additional guidance and information on their use.

References

1. Anvarov, A., Tojaxmedova, I., & Botirova, P. (2015). Learning Resources and Professional Development at Namangan Engineering Pedagogical Institute. YoungScientistUSA, 3(ISBN), 54.
2. Ботирова, П. Х. (2016). Using modular object-oriented dynamic learning environment (Moodle) in NEPI. Молодой ученый, (3), 796-798.
3. Botirova, P., Atamirzayeva, E. B., & Saydaliyeva, M. A. (2019). SPECIFIC FEATURES OF USING INFORMATION TECHNOLOGIES IN LEARNING PROCESS. Theoretical & Applied Science, (5), 634-638.
4. Khakimjonovna, B. P. (2020). Development of coherent speech of students of technical universities in english language education process. European Journal of Research and Reflection in Educational Sciences Vol, 8(11).
5. Botirova, P., & Sobirova, R. (2019). FEATURES OF THE TRANSLATION OF POETRY INTO ENGLISH. Theoretical & Applied Science, (6), 383-387.
6. Botirova, P. (2019). MODERN METHODS OF TEACHING FOREIGN LANGUAGES. Теория и практика современной науки, (2), 25-27.
7. Nargiza, D., & Palina, B. (2019). Features of the english translation of Russian-Speaking realities in the texts of fiction novels. ACADEMICIA: An International Multidisciplinary Research Journal, 9(4), 117-121.
8. Botirova, P. (2019). MODERN PROBLEMS OF LINGUISTICS AND METHODS OF TEACHING ENGLISH LANGUAGE. Теория и практика современной науки, (2), 28-31.
9. Khakimjonovna, B. P. (2021). Methodology of Student Coherent Speech Development in The Process of English Language Learning. Middle European Scientific Bulletin, 9.