The BLARK Matrix and its relation to the language resources situation for the Celtic languages

Delyth Prys
Language Technologies Unit, Canolfan Bedwyr
University of Wales, Bangor, UK
E-mail: d.prys@bangor.ac.uk

Abstract

BLARK (Basic Language Resource Kit) was originally developed as a concept to specify the minimum corpora, tools, and skills needed to engage in pre-competitive research for a language. It was then meant to help identify existing resources, and co-ordinate action to fill any gaps. The BLARK matrix has so far been used for Dutch and Arabic, both fairly well-endowed languages in terms of resources, government support and commercial potential. This paper seeks to explore the usefulness of BLARK for languages in a much weaker position, having very few resources, and little, if any, official support. It examines in particular the six Celtic languages spoken today and asks whether BLARK can be used as a tool to assess their resource needs. It suggests that the lack of basic raw materials, such as the absence of daily newspapers in these languages, is a serious drawback for the development of basic resources, and is not adequately covered in the present BLARK matrix. However, identifying the raw materials needed, and finding alternatives where these do not exist, is in itself a valid exercise. This paper proposes therefore the creation of a preliminary matrix, or PRELARK, to aid the development of resources for these languages.

1. BACKGROUND TO BLARK

The Basic Language Resource Kit or BLARK was originally conceived of by Steven Krauwer and proposed as a cooperative initiative between ELSNET (European Network of Excellence in Language and Speech) and ELRA (European Language Resources Association). The hope was that “with such an action, every European Language, inside or outside the European Union, could have its own BLARK” (Krauwer 1998). This led to the adoption of the BLARK concept for the Dutch language (Cucchiarini, Daelemans & Strik 2001). The BLARK matrices were then developed and published on the web by ELDA (http://www.elda.org/blark/) with an exercise to gather information on Arabic resources from the NEMLAR project (Maegaard 2004) completed. The BLARK website is an interactive one, with an open invitation to all languages to engage with the exercise and fill in the matrices, or send in proposals for a new matrix and other relevant comments.

2. WELSH LANGUAGE RESOURCES

Interest in using the BARK matrices to help the Celtic languages was first expressed at a workshop at the Language Technologies Unit[1], based at Canolfan Bedwyr in the University of Wales, Bangor, in December 2005. This Unit had, for many years been involved in the creation of language resources for Welsh. These include terminology and lexical databases, spelling and grammar checkers, and more recently, speech processing resources. The Unit was, and remains, entirely self-funded, and had therefore undertaken projects on a needs-led basis, responding to invitations to tender where it had the necessary skills to offer, and developing grant proposals where suitable funding opportunities could be identified. However, it was acutely aware of the ad hoc nature of these developments and felt the need for a 'roadmap' or audit of resources available and needed in order to plan future projects in a more comprehensive and orderly manner. The BLARK concept provided it with an off the peg solution to its needs.

BLARK was seen as providing greater detail and more systematic coverage, compared to other attempts to provide a coherent framework for Welsh language technologies. Specifically, it compared favourably to the Welsh Language Board IT Strategy Document, published in March 2006. This document "aspires to the normalisation of the Welsh language in the world of Information Technology". It lists 64 targets and policy statements in order to fulfill its aims, and are relevant to the creation of a digital infrastructure. However, they are conceived of as discrete targets, with the focus on encouraging, discussing, facilitating, engaging in dialogue, examining possibilities etc, towards developing specific applications, such as machine translation, rather than the establishment of a cohesive, coherent infrastructure to support the proposed activities. There is therefore no consideration of the importance of basic resources, such as written and speech corpora, which may be reused and recycled in many applications, thus saving time and money.

3. OTHER CELTIC LANGUAGES

Some of the projects being undertaken by the Language Technologies Unit were of a multilingual nature, involving other Celtic languages. These included two projects funded under the EU Interreg IIIA Wales/Ireland Programme. One was WISPR (Welsh and Irish Speech

[1] Originally founded as the Centre for the Standardization of Welsh Terminology, it subsequently became part of the e-Welsh Unit which has recently been renamed to reflect more accurately its multilingual role. The Canolfan Bedwyr website may be found at http://www.bangor.ac.uk/ar/cb/index.php

Processing Resources), and the other was Lexicelt, an on-line interactive Welsh/Irish Phrasebook and Dictionary. The question was therefore asked whether, if the BLARK matrices could prove useful for Welsh, it could also be extended to the other five Celtic languages. All six Celtic languages share common socio-political situations, as well as a common linguistic structure and heritage (Ó Néill, 2005). With the exception of Irish, which has recently received full status as an official language of the EU, none of the Celtic languages has an official status in the European Union, and all of them, including Irish, have relatively few native-standard speakers[2].

More important than the official status or number of speakers however was the lack of some basic materials from which to create language resources. For example, language corpora of various kinds are core elements for a number of language resources. Written corpora depend on large amounts of text, such as that provided by daily newspapers. However, none of the Celtic languages currently has a daily newspaper, and therefore, the task of collecting material for a basic, balanced corpus is problematic. Material collected from the internet is increasingly being used for the creation of large scale corpora, and software such as Kevin Scannells' *An Crúbadán* is able to crawl the web and compile corpora automatically from web pages in minority languages. It has been argued that using the web as a huge on-line corpus does away with the need for balanced, representitively chosen samples, as everything will be included. While this may work for larger languages with comprehensive coverage of material on the web, it is less satisfactory for very small languages, for example Cornish, where the material may be mainly the work of a very small core of enthusiasts, with, for example, very little 'official' texts. Neither does it distinguish between original and translated material. In the case of Welsh, for example, public bodies have to maintain bilingual websites, but this material almost always originates in English, and the Welsh versions are translations. Translated material may be of poor linguistic standard, and unsuitable for use as a basis for applications such as semantic analysis. One positive aspect of the extensive use of translation in such circumstnaces is that bilingual text corpora at least should be relatively easy to build, paving the way for applications such as machine translation and bilingual lexicons.

Modules such as transcription of broadcast news are also difficult to gather if there is no, or very little, radio and telvision news being broadcast, as is the case with Irish and Manx, and to a lesser degree, Breton and Scots Gaelic. Under such circumstnces new stratergies may need to be devised, such as the recording of live plays or the commissioning of specially prepared scripts to be read aloud and recorded.

The lack of teaching and research into some of these Celtic languages is also detrimental to the creation of basic resources. Cornish and Manx do not have institutes of higher education on their territories, and there are no undergraduate courses fully dedicated to teaching these languages. This leads to lack of basic materials such as contemporary phonetic descriptions, up-to-date grammars, and accurate place-name information.

4. THE NEED FOR A PRELIMINARY BLARK

While BLARK was designed to be language independent, and to be geared towards a "smaller language of Europe" like Dutch, (Krauwer, 2003), it is still geared towards languages which are in a different league from minority languages such as the Celtic ones which are unlikely to receive much private or public funding for language technologies. While Krauwer did envisage the scenario of some languages having to "start from scratch" to create BLARK components, the reality of being faced with a matrix where the score is repeatedly zero as to availability, and high in terms of prioritized need, can be daunting.

What would be helpful therefore, would be a cut-down version of BLARK, or a PRELARK (Preliminary Language Resources Kit) which would identify the very first components needed to provide entry level applications into language technologies. This would focus not only on the intended basic applications envisaged, but also on the limitations placed the lack of available raw materials. While this would prove useful to many smaller minority languages in Europe, the Celtic languages would be well placed to take immediate advantage of it, to build on their existing networks of cooperation, and use it to share and develop common resources in the wider European context.

5. BIBLIOGRAPHICAL REFERENCES

Krauwer, Steven, (2003). The Basic Language Resource Kit as the First Milestone for the Language Resources Roadmap. http://www.elsnet.org/dox/krauwer-specom2003.pdf

LEXICELT Welsh/Irish on-line Phrasebook and Dictionary http://www.lexicelt.org/

Ó Néill, Diarmuid (ed.) (2005). Rebuilding the Celtic Languages. Y Lolfa Press.

Scannell, Kevin P. (2004). Corpus Building for Minority Languages. http://borel.slu.edu/crubadan/

Welsh Language Board (2006). Information Technology and the Welsh Language: A Strategy Document. Cardiff

WISPR (Welsh and Irish Speech Processing Resources) website http://www.bangor.ac.uk/ar/cb/wispr.php

[2] Note however, that the Isle of Man is not a part of the European Union, but a dependency of the British Crown.