Speech and Language Technology on a shoestring and how to get there in a hurry

Michael Bauer
Akerbeltz & iGàidhlig
1/2, 47 Wilton Street, Glasgow G20 6RT, Great Britain
fios@akerbeltz.org

Abstract

Technology permeates our lives more and more. The question of how to tackle this issue is becoming an increasing issue for smaller language communities aiming to conduct their daily affairs through the medium of their language. There are numerous challenges but as we will demonstrate, much can be achieved even with a limited amount of manpower and money available if efforts are planned and focussed.

1. Why bother about Speech and Language Technology (SALT) in small languages?

Small languages often struggle to gain a foothold in Speech and Language Technology.[1] Efforts in this area often end up in disjointed, isolated and uncoordinated projects, in turn resulting in wasted resources and slow progress. Yet with technology increasingly permeating our lives, the need to plant a linguistic flag in cyberspace becomes more important for small and medium languages every year as there is at least some evidence to support the intuitive notion that the metalanguage of technology impacts our wider patterns of language use.[2] [3]

Coupled with these more subtle effects are the more practical issues of how to interface with technology if the technology does not support your language readily – such as predictive texting, auto-correction and voice recognition.

The negative impact of domain loss/gaps, including scientific and technological domains, on the perceived status of a language is also well known[4] [5] and the need for ‘modernizing the image of a language’ frequently appears in attitudinal studies of minority languages.[6]

Ultimately there is most likely a need for a linguistic SALT rights charter at the EU level aimed at private sector multinationals but in the meantime, using the Scottish Gaelic experience as a case study, I will argue that if efforts are focussed, targeted and planned, the mere equivalent of approx. 2.5 FTEs can gain a language a significant digital foothold, i.e. a level of presence allowing everyday users of a language to conduct a large percentage of their daily technological interactions in their language.

2. What to aim for and how to get there on a shoestring

The ultimate goals are to have the maximum number of speakers using a wide range of SALT in their language, producing these with a minimum of resources and within a short time-frame and using future-proofed approaches[7] where possible. For this, three main ingredients are needed:

- A prioritised roadmap for lexicographical and SALT development

[1] Used here in a very loose sense to refer to any interface of language and digital technology, from localized user interfaces and digital wordlists to text to speech or voice recognition tools.

[2] For a slightly more detailed look at this issue, please see Bond & Bauer *GAELIC 2.0 - Advances and New Challenges* in JCLL Volume 17, 2013.

[3] Morris, Cunliffe & Prys *Social Networks and minority language speakers* in SOLS 6.1. 2012

[4] Crystal *Language Death*, Cambridge University Press, 2000

[5] O'Rourke *Galician and Irish in the European Context – Attitudes towards Weak and Strong Minority Languages,* Palgrave Studies in Minority Languages and Communities 2011

[6] MacNeil & Stradling *Emergent Identities and Bilingual Education,* Lèirsinn Research Centre 1999-2000

[7] Streiter, Scannel & Stuflesser *Implementing NLP Projects for Non-Central Languages* in Machine Translation, 2007 (revision)

- A translator (with at least an interest in technology and computing), a lexicographer (or at least someone with a keen interest in dictionaries) and an IT developer. The developers usually do not have to be involved full-time.
- A willingness to cooperate with larger, existing projects, preferably Free and Open-Source Software (FOSS) for better sustainability.

This is not dissimilar to the concept of a BLARK,[8] but less focussed on linguistic corpora which, in their pure form are likely to remain outwith the reach of smaller languages.

3. How does it work?

Gaelic was fortunate as in 2009 Prof Kevin Scannell, an experienced localizer and developer, gave us some sound advice (which was duly heeded) at a pivotal point regarding key aspects of a technological and lexicographical roadmap. But ideally one should try not to rely on luck.

In our case, this led to a radically different design of a planned dictionary project (which would otherwise not have involved a lexical database, see 3.1) and the localization[9] of Mozilla Firefox. In turn, this lead to an almost natural progression of downstream projects.

Using a slightly modified map (the benefit of hindsight), three key elements would appear to be:

- A dictionary project (or at least a very advanced and well-maintained wordlist)
- Localization and development efforts
- Dissemination efforts (incl. boosting user trust in the organisation promoting the tools)

3.1 Dictionaries

Dictionary projects historically waste much effort by creating printed dictionaries based on text documents. While the immediate bonus is the relatively quick creation of a dictionary and a low technological bar, these are not future-proof[10] as they are not easily amended and are hard to convert into digital tools (for example a spell-checker).

Creating a digital dictionary based on a lexical database is an approach which is initially slower but ultimately leads to a more powerful resource. In the Gaelic case, this was achieved by the *Faclair Beag* (AFB) project, which from the end-user perspective functions more or less like any other online dictionary: bidirectional searches, IPA, sound and grammatical, extended information etc.

As so often, most of the black magic happens behind the scenes where the lexicographical data is stored in tables which themselves are stored as relational databases not only identifying general POS data but also micro-level morpho(phono)logical data, gender, case, tense, mood and person.

The immediate benefit are smart dictionary searches, which means a user can enter an inflected form and be directed to the appropriate root. Experience with AFB has shown this to be highly useful and popular not only for direct users of AFB but also for 3rd party projects such as Wordlink,[11] a project which offers language learners a split browser screen with learning content on the left and via left-clicks an automatic lookup in a given dictionary on the right.

To ease editing work, the editor does not have to work in the back-end but has an easy to use web interface. Editors are furthermore aided in the generation of this data by a feature which automatically generates the various forms of a lexeme. In the case of regular nouns, verbs and adjectives, this means the editor often only has to check but not edit each item, resulting in massive time savings when one considers that the average regular verb has just under 50 forms.

[8] Basic Language Resource Kit, see Krauwer *The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap* in Proceedings of SPECOM2003

[9] Localization refers to the translation and adaptation of software to a specific language and region.

[10] For slightly different reasons, the ‘app trap’ should also be avoided unless significant resources are at hand. App-based dictionaries sound great but involve a lot of time (=money) in terms of development for multiple platforms and almost interminable bugfixing. With the increasing popularity and sinking cost of data services on mobile devices, web-based dictionaries often are a more sustainable resource.

[11] http://multidict.net/wordlink/

This setup allows for substantial growth through the amalgamation of existing sources. In our case, in the 5 years since its launch the total number of lexemes has grown to 34,000 headwords, resulting in 184,000 forms and 877,000 items in total (once affixes like emphatic endings etc. are factored in).[12]

3.2 The Roadmap

The initial stages of the proposed roadmap likely apply to most languages in question but once a certain level is achieved, the exact order of projects can be varied based on a needs analysis, community feedback or indeed requests.

Once the initial projects (the dictionary, browser and office suite) have been tackled, the lexicography and localization/development start feeding off each other in various ways.

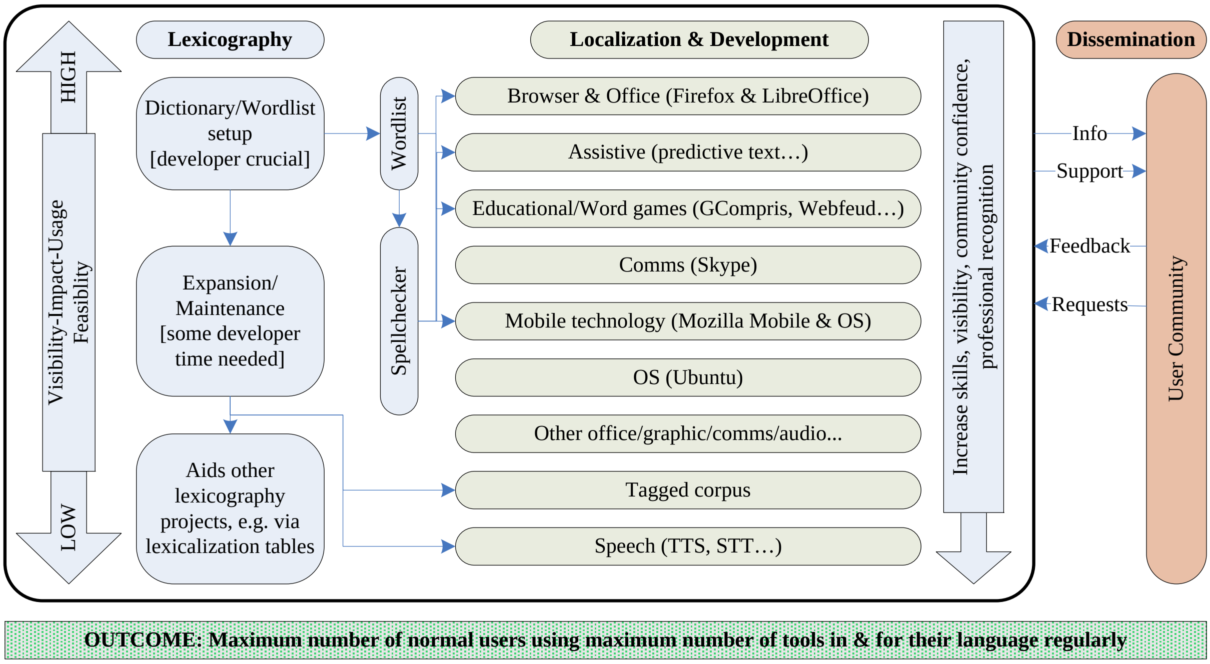
HIGH
Visibility-Impact-Usage Feasiblity
LOW
Lexicography
Dictionary/Wordlist setup [developer crucial]
Expansion/ Maintenance [some developer time needed]
Aids other lexicography projects, e.g. via lexicalization tables
Wordlist
Spellchecker
Localization & Development
Browser & Office (Firefox & LibreOffice)
Assistive (predictive text…)
Educational/Word games (GCompris, Webfeud…)
Comms (Skype)
Mobile technology (Mozilla Mobile & OS)
OS (Ubuntu)
Other office/graphic/comms/audio...
Tagged corpus
Speech (TTS, STT…)
Increase skills, visibility, community confidence, professional recognition
Dissemination
Info
Support
Feedback
Requests
User Community
OUTCOME: Maximum number of normal users using maximum number of tools in & for their language regularly

Figure 1 A proposed roadmap for the most effective path towards a full range of SALT

Developer input is usually not required on a full-time basis but rather at certain junctures (for example an introduction to placeholders, plural formatting, the initial creation of the spellchecker, generating web-based statistics etc.).

3.3 Spreading the word

Details about availability, installation and common issues is provided via the web (www.iGaidhlig.net) and various Social Media platforms. Some face-to-face workshop trials have also been held.

3.4 Common problems

There are various pitfalls, not all of which can be avoided:

- Participating in a commercial or larger FOSS project is often a high-risk strategy that may lead both to great benefits but little guarantee for long-term sustainability. For example the popular Google In Your Language project was ultimately shelved by Google. On the other hand, joining Adaptxt[13] has enabled availability and maintenance on an industry-standard tool. Trial versions of Gaelic Text-to-speech in collaboration with Cereproc[14] are also encouraging. If efforts are to some extent user-demand driven, engaging with such projects may occasionally be unavoidable or even desirable to reduce development workload.
- Especially smaller but even bigger FOSS projects can ‘die’ when key people leave

[12] Data directly from AFB.

[13] A predictive texting app for Android (with a planned release on iOS), www.adaptxt.com.

[14] An Edinburgh-based text to speech company, www.cereproc.com.

- Choosing ‘sexy’ projects which turn out to be high-cost and low-impact such as machine translation (which, if poorly done, can lead to mass production of poor translation) or language-specific social networks (which usually fail due to the assumption that users might be willing to engage in monolingual social networking)
- Technological problems (issues with implementing correct plurals, force locale[15])

4. Evaluating the Gaelic experience to date

The first digital Gaelic tool – the *Stòr-dàta*, an online termbase – appeared in 1994. Between 1994 and 2008, about a dozen other tools appeared, most of which then fell dormant for a period (the Opera web browser, OpenOffice.org and the Ning-based social network *AbairThusa*) or died off when funding/support ran out or the localizer moved on. Since the end of 2009 however, just over 50 additional programs and tools have appeared, ranging from games and web-apps through predictive texting tools to operating systems (Ubuntu, Windows and the upcoming Mozilla OS), allowing users to conduct a large percentage of their daily IT through the medium of Gaelic.

These were almost all created by two (largely) unpaid part-time localizers and two (largely) unpaid part-time developers. Their time involvement is difficult to quantify but an estimate puts it at 1.5 FTE of localizer and lexicographer time and 0.5 of a FTA of developer time over the last four years.

An unexpected benefit of having a small team produce a large number of localizations is unusually high consistency of the translations, especially in terms of terminology. Many FOSS projects suffer from having too many ‘cooks’ spoiling the broth of consistency, within and across projects.

The most significant challenge though is not technical but human. Most everyday users of technology use it ‘as it comes out of the box’ and generally are reluctant to tinker with it unless coached by someone experienced. For example, although use of the Gaelic Firefox is slowly growing, it has taken almost 3 years to grow the userbase by approximately 20 to around 120 regular users. This pattern of low uptake (below what might be expected based on a product’s market penetration) appears to be common across languages (the Irish Firefox has about 300 regular users[16]) and other projects.

Uptake of tools which are for but not in Gaelic is higher (for example, since 2009 AFB has approx. 140,000 searches per month and the predictive texting tool Adaptxt had been download 2,871 times in Gaelic and 3,810 times in Irish by 31 Jan 2014[17]). While this is encouraging, uptake remains an issue. Regarding home users, face-to-face pilot workshops where users are guided through available tools and the installation process have proven popular and the current aim is to expand this, ideally through hiring a peripatetic community ‘promoter’ who would hold free workshops across Scotland.

In spite of interest from the educational and public sector in Scotland, all such efforts are hampered by the current IT provision model. Outside suppliers are contracted to provide (often thin or dumb) clients with limited or no admin rights for the end-users. At best the user may install software not on the approved list on a local system but not across, for example, all computers of a school. Until there is a requirement to provide Gaelic IT alongside the English (or until an alternate route is found), there is little that can be done to improve the provision of Gaelic technology in spite of availability.

5. Our take-home message

- Dissemination of information, user support and promotion must be considered at an early stage, as such tools will not simply disseminate through their mere existence
- FOSS is harder to ‘sell’ to everyday users but ultimately the only really sustainable model for small and medium languages in most cases
- It is nonetheless very doable, as since 2009 Gaelic has acquired a lot of new SALT through the work of small group of people and any language development agency should seriously consider supporting or setting up such a group

[15] A feature which forces the interface language to match that of the operating system (OS). This works well if the OS is available in a given language and if the user is monolingual. Hence, a major impediment in smaller languages.

[16] Weekly data provided by Mozilla to locale leaders.

[17] Data provided to the author by Adaptxt upon request.